UNDERSTANDING HUMAN DEMOGRAPHY AND ITS IMPLICATIONS FOR THE DETECTION AND DYNAMICS OF NATURAL SELECTION

A Dissertation

Presented to the Faculty of the Graduate School

of Cornell University

In Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

by

Kirk Edward Lohmueller

February 2010

UNDERSTANDING HUMAN DEMOGRAPHY AND ITS IMPLICATIONS FOR THE DETECTION AND DYNAMICS OF NATURAL SELECTION

Kirk Edward Lohmueller, Ph.D.

Cornell University 2010

Patterns of genetic variation from living people can provide important information regarding ancient demographic history as well as how recombination and natural selection operate in different populations. However, extracting important information from large-scale datasets remains challenging. In this dissertation, I develop and validate statistical methods to understand human demography, patterns of recombination across the genome, how demography impacts the ability to detect positive selection, and how demography influences the proportion of deleterious genetic variation within a population. First, I develop a novel haplotype-based approach to estimate bottleneck parameters from human single-nucleotide polymorphism (SNP) data. Application of my method to simulated data shows it can reliably infer parameters from growth and bottleneck models, even in the presence of recombination hotspots when properly modeled. Application of the method to data collected by Perelgen Sciences shows evidence for a severe population size reduction in northwestern Europe starting 32,500- 47,500 years ago. Second, I compare patterns of linkage disequilibrium (LD) in HapMap data on the human autosomes to patterns to the human X chromosome. I find too little LD on the X chromosome relative to what is predicted under simple models based on the amount of autosomal LD. Third, I

assess the effect of recent admixture on population genetic methods to infer ancient population growth. Haplotype methods are sensitive to recent admixture while methods based on SNP frequencies are less sensitive. Fourth, I evaluate the effect that recent admixture has on the ability to detect positive selection. Simulations show that admixture causes a decrease in power of some tests of selection, while increasing power for others. These results have important implications for detecting selective sweeps in admixed populations. Finally, I show that demographic history has had an important impact on patterns of segregating deleterious polymorphism in different populations. In particular, exon-resequencing data collected by Celera Genomics shows that European populations have a higher proportion of damaging mutations than African populations do. Through the use of forward-simulations with realistic demographic and selection parameters, I demonstrate that this pattern can be explained by the differing demographic histories of the two populations.

BIOGRAPHICAL SKETCH

Kirk was born in Methuen, MA in 1982 to Barbara and Karl Lohmueller. Kirk's curiosity with the world around him and love of learning came at an early age due to the influence of his parents. He especially enjoyed learning from his dad about home improvements, how household things worked, and why they sometimes did not. Due to watching the O.J. Simpson trial, while in junior high, Kirk became fascinated with forensic science and DNA. Thanks to a series of fortuitous events, Kirk began working at the Whitehead Center for Genome Research while attending Central Catholic High School in Lawrence, MA. While at Whitehead, thanks to the spirited and guiding influence of David Altshuler and Joel Hirschhorn, Kirk's interests switched from forensic genetics to human population genetics, where they remain today. Kirk then attended college at Georgetown University where he continued his research in human population genetics working with Lee-Jun Wong and John Braverman. While at Georgetown, Kirk also met his future wife, Charlene. After college, Kirk began graduate school at Cornell University in the Genetics & Development Field. At Cornell, Kirk received training in probability and statistics and computer programming, topics he did not enjoy very much earlier in his education. While working in the labs of Andy Clark and Carlos Bustamante, Kirk completed many research projects in human population genetics. In 2008, Kirk and Charlene were blessed with a precocious little boy, Walter Alan Lohmueller. With Walter's help, Kirk completed his dissertation research within the next year and half. After graduating from Cornell, Kirk and his family will be moving to Berkeley, CA where Kirk will undertake postdoctoral research with Rasmus Nielsen.

To Wally and Charlene

ACKNOWLEDGMENTS

Special thanks goes to my advisors, Andy Clark and Carlos Bustamante for their continuous support throughout my time at Cornell. Both have given me tremendous freedom and allowed me to pursue whichever research projects I found interesting. Early on during my time at Cornell, they pushed me to learn more math and statistics than I thought I could possibly handle. In retrospect, this has helped me tremendously, and so I wish to thank them. I would also like to thank my committee members, Andy Clark, Carlos Bustamante, Brian Lazzaro, and Jason Mezey for their time and help, and especially Brian Lazzaro for his helpful comments on this dissertation. Special thanks goes to Jeremiah Degenhardt for helping me think through much of the research presented here. Without Jeremiah's help, these ideas would not be as developed or carefully argued. The wonderful friends I have made at Cornell, especially Adam Boyko, Kasia Bryc, Andy Reynolds, Adam Auton, Amit Indap, Kirsten Eilerston, Melanie Huntley, Amanda Larracuente, Ryan Hernandez, Dara Torgesson, and Jeremiah Degenhardt, have made my time here as enjoyable as it was—thank you. Mom and Dad, thank you for all that you have taught me and sacrificed for me to have reached this point. Thank you to my wonderful wife, Charlene, for her constant support, encouragement, and patience, when I had to be working instead of helping out with Wally and the household chores. Her organization, culinary skills, and generosity have prevented my research from slowing down after Wally's arrival. Lastly, thank you Wally for your patience when "DaDa" had to be "playing" on the computer instead of playing with you.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

PREFACE

A major challenge to the study of population genetics is that it is often a historical science, rather than an experimental one. For example, human evolution occurred only once. It cannot be carefully repeated again in a laboratory. As such, all of population genetics relies on some sort of model, either implicitly or explicitly, to test hypotheses and to help interpret patterns of genetic variation from extant populations (reviewed in Marjoram and Tavaré 2006). Models are used to test whether a particular evolutionary scenario is likely to have given rise to the observed patterns of genetic variation that are seen in empirical data. In this dissertation, I use population genetic models to learn about human demographic history and natural selection.

One area of great research interest in population genetics is the use of genetic variation data to infer population demographic history. Again, this can be done either assuming an explicit demographic model, or can be done by qualitatively comparing statistics across populations. Explicitly assuming a particular model framework allows one to infer which parameter combinations are most consistent with the observed data (reviewed in Nielsen and Beaumont 2009). Explicit models can also be tested to determine whether or not they are compatible with the data.

The method that is the current state-of-the art for inferring demographic parameters from genetic variation summarizes the single nucleotide polymorphism (SNP) data by the site frequency spectrum (SFS; Nielsen 2000; Adams and Hudson 2004; Williamson *et al.* 2005; Boyko *et al.* 2008; Gutenkunst *et al.* 2009). The SFS is simply the count, or proportion, of SNPs at a particular frequency in the sample. The SFS is sensitive to various demographic and selective factors (reviewed in Wakeley 2008). Either coalescent simulations (Nielsen 2000; Adams and Hudson 2004) or

diffusion theory (Williamson *et al*. 2005; Boyko *et al*. 2008; Gutenkunst *et al*. 2009) is used to obtain the expected SFS for a given set of demographic parameters and then the model parameters that best match the SFS from the observed data are chosen. There are two major criticisms of the SFS approach: 1) it is difficult to apply it to SNP data that have not been discovered through direct resequencing of all the individuals in a sample (Nielsen *et al*. 2004; Clark *et al*. 2005) and 2) it ignores the linkage disequilibrium (LD), or correlations, among the SNPs (Myers *et al*. 2008).

To address these concerns with the SFS-based approaches to demographic inference, in Chapter 1 of my dissertation, I propose a novel haplotype-based method to infer demographic history, called the Haplotype Count-Number (HCN) approach. I perform extensive simulations to evaluate the method's performance under many different scenarios. I also illustrate its utility by applying it to data and estimating bottleneck parameters for a European population. Finally, I discuss how haplotype and LD information may be more informative than the SFS-based approaches at distinguishing complicated demographic scenarios.

The study presented in Chapter 1 analyzed autosomal patterns of haplotype variation. However, due to the fact that males carry only one X chromosomes, while females have two, comparing patterns of genetic variation between the autosomes and the X chromosome can reveal important evolutionary insights. For example, under the simplest neutral models, it is predicted that the effective population size of the X chromosome should be ¾ that of the autosomes (reviewed in Hedrick 2007). Two recent studies reported departures from this ratio in human populations and suggested sex-biased demographic processes as an explanation (Hammer *et al*. 2008; Keinan *et al*. 2009). Since neither study analyzed LD patterns, in Chapter 2, I compare patterns of LD on the X chromosome and the autosomes using HapMap data (International HapMap Consortium 2007). This chapter illustrates how LD patterns can be affected

by both the demographic and recombination processes, making it difficult to disentangle the effects of these two evolutionary processes.

The first two Chapters of my dissertation make extensive use of demographic models to interpret patterns of LD. A major criticism of population genetic models is that they are too simple and fail to capture all the intricacies that really occurred throughout evolutionary history (Marjoram and Tavaré 2006; Nielsen and Beaumont 2009; Templeton 2009). The hope is that simple models capture enough of the important aspects of the evolutionary process and that the omitted details do not really impact the observed patterns of variation. As George Box said, "all models are wrong, some are useful" (Box 1979). To determine whether details that are often not considered in modeling studies are actually influence the results, it is useful to simulate data containing the details and then determine what impact the details have on patterns of variation. Additionally, statistical tests and methods of inference that ignore the details can be evaluated for their accuracy. In Chapters 3 and 4 of this dissertation, I explore the effect of an important demographic detail that is often omitted from studies of human demography and selection—recent admixture. In particular, African Americans derive from a recent mixture of European and African ancestry (see for example Tishkoff *et al.* 2009), but population genetic studies often treat them as being purely African (Adams and Hudson 2004; Akey *et al.* 2004; Marth *et al.* 2004; Carlson *et al.* 2005; Kelley *et al.* 2006; Tang *et al.* 2007; Williamson *et al.* 2007; Boyko *et al.* 2008). In Chapter 3 I evaluate whether the SFS and HCN methods of demographic inference are affected by the recent admixture. In chapter 4, I evaluate whether commonly used tests of neutrality exhibit false positive results when applied to admixed populations. I also evaluate whether there is a loss in power to detect selective sweeps when using individuals from an admixed populations. Both of these studies are critically important to interpreting previous studies of human

demography and directional selection.

Another important violation of common models of weak negative selection is changes in population size. Traditional predictions of the SFS of mutations under weak selection were made assuming constant population size (Wright 1938; Kimura 1964; Sawyer and Hartl 1992; Hartl *et al.* 1994). These models have recently been extended to include the case of changing populations size (Williamson *et al.* 2005; Boyko *et al.* 2008). In Chapter 5 of my dissertation, I demonstrate the practical relevance of such models. I examine patterns of synonymous and nonsynonymous SNPs in resequencing data from European and African American individuals. I show that there is a difference in the proportion of SNPs that are nonsynonymous, and are likely to be deleterious, between the two populations. This chapter illustrates the importance of modeling demographic history when thinking about weak/moderate negative selection.

CHAPTER 1

METHODS FOR HUMAN DEMOGRAPHIC INFERENCE USING HAPLOTYPE PATTERNS FROM GENOME-WIDE SINGLE NUCLEOTIDE POLYMORPHISM DATA[1]

1.1 Abstract

We propose a novel approximate likelihood method to fit demographic models to human genome-wide single nucleotide polymorphism (SNP) data. We divide the genome into windows of constant genetic map width, then tabulate the number of distinct haplotypes and the frequency of the most common haplotype for each window. We summarize the data by the genome-wide joint distribution of these two statistics—termed the *HCN* statistic. Coalescent simulations are used to generate the expected *HCN* statistic for different demographic parameters. The *HCN* statistic provides additional information for disentangling complex demography beyond statistics based on single-SNP frequencies. Application of our method to simulated data shows it can reliably infer parameters from growth and bottleneck models, even in the presence of recombination hotspots when properly modeled. We also examined how practical problems with genome-wide datasets, such as errors in the genetic map, haplotype phase uncertainty, and SNP ascertainment bias, affect our method. Several modifications of our method served to make it robust to these problems. We have applied our method to data collected by Perelgen Sciences and find evidence for a severe population size reduction in northwestern Europe starting 32,500-47,500 years ago.

---

[1] Previously published in Lohmueller *et al.* (2009) and has been reproduced with permission. Copyright Genetics Society of America.

1.2 Introduction

A major goal of evolutionary genetics is to infer the demographic history of a population. This is traditionally done by fitting a population genetic model to sequence data taken from a sample of individuals. The population genetic model often includes parameters allowing for changes in population size or population structure with or without migration. Such parameters are interesting in their own right, but are critical to define a proper "null model" which can be used to find "unusual" genes that may be targets of positive or negative selection (Jensen *et al.* 2005). Additionally, a proper demographic model is important for assessing genome-wide patterns of positive and negative selection (Boyko *et al.* 2008; Lohmueller *et al.* 2008).

Methods have been developed that make full use of sequence data to infer demographic parameters (Griffiths and Tavaré 1994; Kuhner *et al.* 1995). These methods are computationally intensive and are impractical for all but the smallest datasets. Thus, researchers have turned to methods based on summary statistics (reviewed in Marjoram and Tavaré 2006). Summary statistics can be quickly calculated from the data and then be used to infer model parameters using either a likelihood or approximate Bayesian computation (ABC) framework (for example Wall 2000a; Fagundes *et al.* 2007). The key for successful application of this approach is to find summaries of the data that contain enough information about the demographic parameters of interest. One of the most successfully used summary statistics for population genetic inference, the site frequency spectrum (SFS; Nielsen 2000; Adams and Hudson 2004; Caicedo *et al.* 2007; Hernandez *et al.* 2007b), is a sufficient statistic for the full data if the SNPs are unlinked. However, in reality, all SNPs are not unlinked. The amount of information in the data lost by ignoring the correlations among SNPs, or linkage disequilibrium (LD), in demographic inference is an open

question, but recent theoretical work suggests that it may be non-negligible (Myers *et al.* 2008).

An additional complication to using the SFS for demographic inference is that many genome-wide genetic variation datasets in humans contain SNPs that were discovered through sequencing a small number of individuals. The discovered SNPs were then genotyped in a larger set of individuals, sometimes in a different population than was used for SNP discovery. Since this SNP discovery process will lead to preferential sampling of intermediate-frequency alleles, the SFS computed from SNP genotype data will differ substantially from the true SFS (Nielsen *et al.* 2004; Clark *et al.* 2005). Progress has been made to analytically correct the SFS for ascertainment bias when the SNP discovery process is known (Nielsen *et al.* 2004), but often this is not the case. More problematic is the situation where SNPs were discovered by resequencing individuals in one population, but then are genotyped in a second population. It remains an open question as to how well the SNPs discovered in the first population are representative of genetic diversity in the second population. Several authors have suggested that statistics based on combinations of multiple SNPs, or haplotypes, will be less susceptible to ascertainment bias than single-SNP frequencies or heterozygosities (Conrad *et al.* 2006). However, while this suggestion is encouraging, as yet there has not been extensive investigation into the precise ascertainment conditions under which this is true.

It is known that haplotype patterns and LD can be affected by both recombination and demographic history (Pritchard and Przeworski 2001), making these measures useful statistics for inference. Many recent studies have assumed a demographic model (often the standard neutral model) and then used either LD or haplotype patterns to estimate recombination rates (Wall 2000a; Hudson 2001; Li and Stephens 2003; McVean *et al.* 2004a; Myers *et al.* 2005). Other studies have taken the

opposite approach and assumed that the recombination rate is known and then used LD or haplotype patterns to estimate demographic parameters (Reich *et al.* 2001; Innan *et al.* 2005; Schaffner *et al.* 2005; Voight *et al.* 2005; Leblois and Slatkin 2007; Tenesa *et al.* 2007). The way in which haplotype information has been used for demographic inference is quite variable among studies. For example Reich *et al.* (2001) examined how well several different demographic models predicted the observed decay of pairwise LD in humans, rather than estimating the model parameters. Thornton and Andolfatto (2006) and Francois *et al.* (2008) used ABC to estimate model parameters in *Drosophila* and *Arabidopsis*, respectively. However, summaries based on the distribution of the number of haplotypes was only one of several summary statistics considered, and it is unclear how much information came from the haplotype information versus the other single-SNP diversity measures. While Anderson and Slatkin (2007) and Leblois and Slatkin (2007) developed methods that use haplotype information exclusively to fit a population split followed by growth model, their model is quite restrictive and only allows inference of one free parameter, the number of founding lineages. Thus, there has not been a systematic investigation as to the utility of haplotype information for inference in general, parameter rich models, such as those involving population expansions and bottlenecks.

In this article we propose an approximate likelihood method to estimate parameters in complex demographic models from genome-wide SNP genotype (rather than full re-sequencing) data using the joint distribution of the number of haplotypes and frequency of the most common haplotype in windows across the genome. We provide extensive simulations evaluating the performance of our method for growth and bottleneck models. These results indicate that a great deal of information regarding demographic history is captured by these two summary statistics. We also extensively test the robustness of our method to many practical problems with genetic

datasets in humans. Specifically, we show that for many realistic SNP discovery protocols and levels of population divergence, our method is relatively robust to SNP ascertainment bias. We also found that our method is sensitive to recombination rate variation across the genome (as many haplotype-based summaries will be), and we incorporate a model of recombination rate variation into the inference scheme. Finally, since haplotype phase is often ambiguous, we provide a practical approach to circumvent this problem. We applied our method to genome-wide SNP genotype data generated by Perlegen Sciences (Hinds *et al.* 2005). Using the CEU sample (consisting of individuals from Utah with Northwestern European ancestry), we find evidence for a recent population bottleneck in Northwestern Europe.

1.3 Methods

Summary statistics

We summarize the genome-wide data by the joint distribution of two haplotype statistics calculated from windows across the genome. Our method requires that we have a genetic map of the organism in question. Using this map, we divide the genome into windows of fixed genetic map distance, c_{window}. The parameter c_{window} is tunable to the diversity and recombination rates of the organism under study. We chose to divide the genome into n_{window} non-overlapping windows using genetic map distance so that each window will have the same expected amount of recombination within it, and consequently, the same expected number of haplotypes (Wall 2000a).

In many genome-wide SNP datasets, some parts of the genome will have a small number of SNPs while other areas will contain many SNPs. In principle, while this could be due to mutation rate variation across the genome, variations in the time to the most recent common ancestor, or random chance, another likely explanation is ascertainment bias—some parts of the genome were more extensively screened for SNPs than others and consequently have more SNPs. Thus, we do not want our

method to use any information about the number of SNPs within a given window. To ensure that all windows of the genome have the same number of SNPs, we select a sub-set of n_{snp} SNPs for each window of the genome. Again, n_{snp} is a function of the size of the windows as well as the SNP density. Another complicating factor is that SNPs may not have been discovered from the population under study, but instead from a second population. Since rare SNPs are more likely to be population specific, and consequently not equally ascertained in all populations, we only include those SNPs with minor allele frequency (MAF) $\geq 10\%$.

Having selected a sub-set of intermediate frequency SNPs from each window, we can compute the number of distinct haplotypes as well as the count of the most common haplotype in a sample of *n* chromosomes. The *HCN* statistic is the genome-wide joint distribution of these two statistics. Specifically, let $X = (X_{1,1}, X_{1,2}, .., X_{1,l}, X_{2,1}, .., X_{2,l}, .., X_{k,1}, .., X_{k,l})$, where X_{ij} denotes the number of windows having *i* haplotypes where the most common haplotype has count *j* out of *n*. In principle $k = l = n$, however, in practice, we bin intervals in the *HCN* statistic for the inference so that fewer simulation replicates will be needed to obtain an accurate estimate of the expected *HCN* (see below), and thus there are fewer than n^2 bins in the *HCN* statistic. Ideally, we would like to integrate over all possible sets of n_{snp} SNPs within each window when constructing the *HCN* statistic. However, this is not computationally feasible, so we generate 10 random matrices (*X*), each using a different randomly selected set of n_{snp} SNPs from each window. We then average these 10 matrices as our final *X* matrix to be used for inference. This is done to reduce Monte Carlo variance resulting from selecting a single set of SNPs. An example of the *HCN* statistic for several demographic models is shown in Figure 1.1.

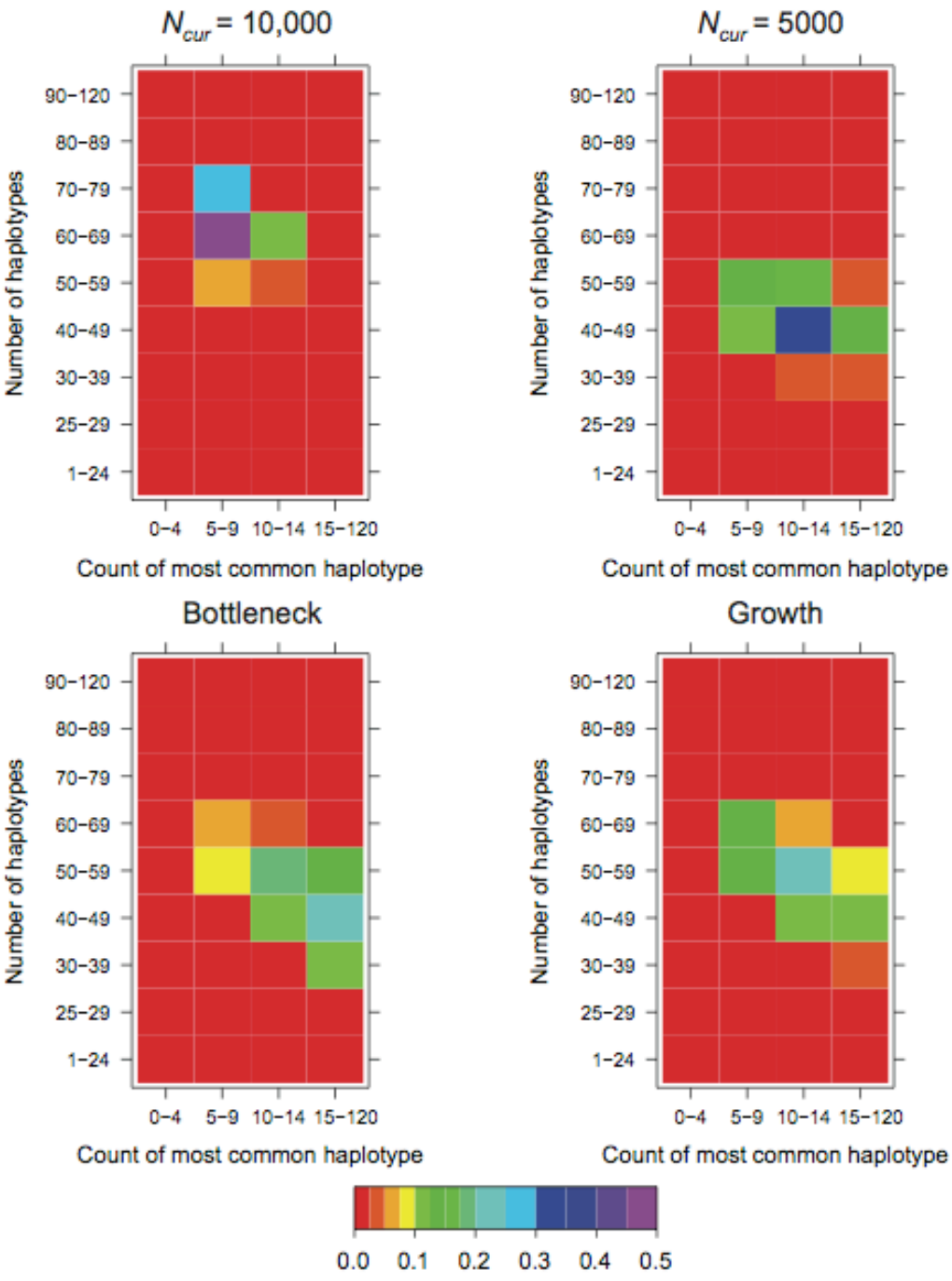


Figure 1.1: Examples of the *HCN* statistic for different demographic models. The color of each cell in a matrix denotes the proportion of simulated windows having the particular number of haplotypes and frequency of the most common haplotype. Approximately 3000 windows were simulated for each demographic model with n_{snps} = 25, c_{widnow} = 0.25 cM. The parameters for the bottleneck model are N_{cur} =N_{anc}= 10,000; N_{mid}=1000; t_{cur}=t_{mid}=800 generations. The parameters for the growth model are N_{cur}=10,000; N_{mid}=1000; t_{cur} =800 generations.

We chose to use the number of haplotypes as a summary statistic because it is a sufficient statistic for the population mutation rate (θ) in the infinite alleles model (Ewens 1972) and has been shown by simulation to be informative about population history (Depaulis and Veuille 1998; Innan *et al.* 2005). The count of the most common haplotype was also suggested as a test statistic in the infinite alleles model (Ewens 1973) and has been found to be correlated with haplotype homozygosity (Zeng *et al.* 2007 and data not shown). The joint distribution of these two statistics performs better at distinguishing among demographic models that using either summary on its own (Figure 1.1). For example, the number of haplotypes is more informative about overall population size than is the count of the most common haplotype (compare N_{cur}=10,000 to N_{cur}=5000), as expected, since larger populations have a higher population recombination rate, $\rho = 4N_e c$ than smaller populations, resulting in a larger number of haplotypes per window (Wall 2000a). Note that because we selected n_{snp} SNPs with $\text{MAF} \geq 10\%$ per window, the fact that the larger population has a higher value of θ does not inflate the observed number of haplotypes per window. A recent bottleneck results in an intermediate number of haplotypes, but the stronger signature of the bottleneck is the excess proportion of windows where the most common haplotype is at unusually high frequency. These patterns are due to an elevated rate of coalescence during the bottleneck, which, for some simulated windows, results in there being fewer lineages available to recombine. A recent population expansion also results in an intermediate number of haplotypes, but without an increase in the number of windows where the frequency of the most common haplotype is unusually high.

The *HCN* statistic contains no information about how different haplotypes within a window are from each other. To add this information, we also considered another summary statistic H_{pair}, the distribution across the genome of the average

number of pairwise differences between haplotypes. For all pairs of haplotypes within a given window, we simply computed the number of SNPs (which could range from 0 to n_{snp}) where the two haplotypes differed and compute the average. H_{pair} is the vector giving the number of windows having a given number of average pairwise differences. We show (APPENDIX 1 and Figure 1.2) that this statistic is not robust to SNP ascertainment bias and do not use it in further analyses.

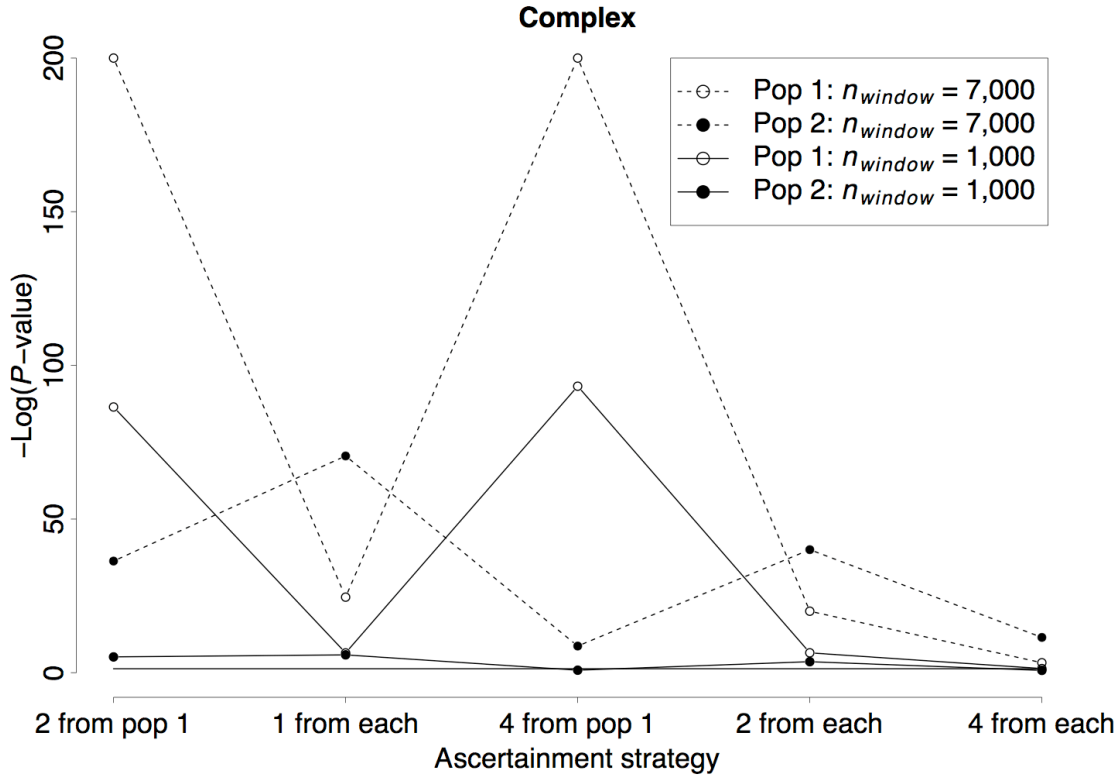


Figure 1.2: Log_{10} *P*-value of the goodness-of-fit test comparing the H_{pair} statistic under different SNP ascertainment schemes (shown on the *x*-axis) to that with complete ascertainment for the complex demographic model. Here a sample size of 40 chromosomes from each population is used. The solid horizontal line denotes the 5% significance cutoff. *P*-values $<10^{-200}$ are set to 10^{-200}.

Demographic models

We consider two different single-population demographic models. These models and their associated parameters are shown in Figure 1.3. Figure 1.3A shows a 2-epoch model which is used for modeling population growth. Here there are three

parameters to estimate: the current population size, N_{cur}, the ancestral population size, N_{mid}, and the time that growth has occurred, t_{cur}. Figure 1.3B shows a three-epoch model which has five free parameters: the current population size, N_{cur}, the population size during the bottleneck, N_{mid}, the ancestral population size, N_{anc}, the time when the bottleneck started (going backwards in time), t_{cur}, and the duration of the bottleneck, t_{mid}. All times are in units of generations. We note that although these models (and all models in population genetics) are arbitrary simplifications of the true demographic history, the hope is that they capture some essential features of population history.

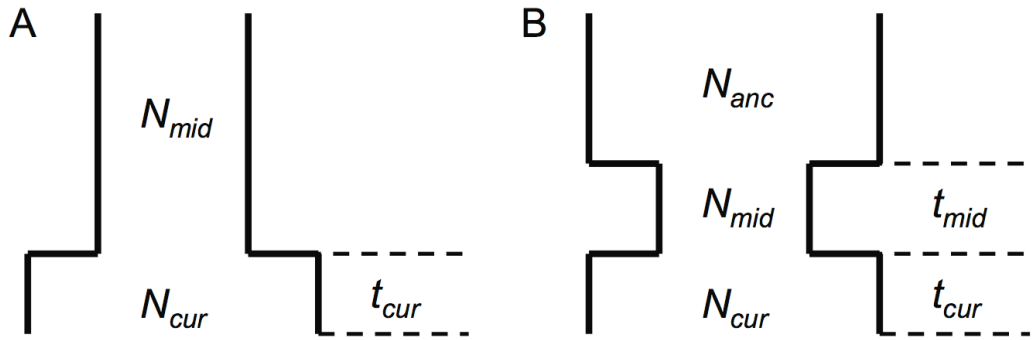


Figure 1.3: Demographic models considered. Relevant free-parameters are shown on each diagram. A) Two-epoch and B) Three-epoch models.

Fitting models to data

Since the observed *HCN* statistic follows a multinomial distribution, we fit demographic models to the data using an approximate likelihood approach (see Weiss and von Haeseler 1998; Wall 2000a; Fearnhead and Donnelly 2002; Plagnol and Wall 2006). We define $\mathbf{p} = (p_{1,1}, p_{1,2}, .., p_{1,l}, p_{2,1}, .., p_{2,l}, .., p_{k,1}, .., p_{k,l})$, where p_{ij} is the probability that a window has i haplotypes where the most common haplotype is at count j. The approximate likelihood function for the demographic parameters (Θ) can be written as

$$L(\Theta) \approx \prod_{i=1}^{k} \prod_{j=1}^{l} p_{ij}^{X_{ij}} \qquad (1.1).$$

We use the coalescent with recombination (Hudson 1983; Hudson 2002) to find p_{ij} for the demographic parameter combination of interest. We simulate z replicates using the demographic parameters of interest (Θ) and $\rho = 4N_{cur}c_{window}$. We estimate the matrix $\mathbf{p}$ as the proportion of simulation replicates falling in a particular bin of the *HCN* statistic. Formally, define the indicator function $I(w,i,j)$ to be equal to 1 if simulation replicate w has i haplotypes and the count of the most common haplotype is j, and equal to 0 otherwise. Then

$$p_{ij} = \frac{\sum_{w=1}^{z} I(w,i,j)}{z} \qquad (1.2).$$

Since we select n_{snp} SNPs from each window, θ does not explicitly enter into these simulations. Therefore, instead of setting an arbitrary value of θ, and then randomly selecting n_{snp} SNPs, we use the "fixed S approach" (Hudson 1993) to add mutations onto the ancestral recombination graph (ARG). Specifically, n_{snp} mutations are randomly placed onto each simulated ARG such that these SNPs will have MAF $\geq 10\%$. To reduce Monte Carlo error, this process is repeated 10 times for each ARG. Each time, we evaluate $I(w,i,j)$ and increment the appropriate bin of $\mathbf{p}$. Note, we record 10 different $\mathbf{p}$ matrices and after the desired number of simulation replicates, we keep the average of the 10 $\mathbf{p}$ matrices as our final matrix. This is an approximate likelihood function since we are approximating $\mathbf{p}$ using simulations, rather than calculating them exactly, and we also treat all windows of the genome as being mutually independent.

We optimize the likelihood function described above using a grid search since we are approximating the likelihoods by simulation and the simulation variance may be non-negligible, misleading deterministic hill-climbing approaches. The number of

grid points and number of simulation replicates used to maximize the likelihood function varies among analyses and are given below.

Since variation in recombination rates at a fine scale can affect the *HCN* statistic, we have added a model of recombination hotspots into our inference method. We describe the parameters used for specific instances below. Since each window of the genome corresponds to the same genetic map distance (c_{window}), the number of base pairs per window will differ among windows. In our simulations to find $\mathbf{p}$, we select the size of the window in base pairs (denoted L) from the observed distribution of physical distance. We then set r, the per base-pair recombination rate, to be constant across the window such that rL will give c_{window}. We then simulate an ARG in the normal manner with recombination rate rL, but then, similar to the method used by Li and Stephens (2003), we model hotspots by changing the relationship between physical and genetic distance. Informally, the parts of the window where hotspots occur are assigned fewer base pairs, and consequently have a lower probability of a SNP occurring in them, than windows with lower recombination rates.

Simulations to evaluate performance

We tested the performance of our method by simulating data under three different demographic models: 1) ancient population growth, 2) recent population growth, and 3) a recent population bottleneck. The parameter values for these models are shown in Figures 1.4-1.7. These models were chosen because of their relevance to human demographic history. For each model we simulated datasets (500 for models assuming uniform recombination rates and 100 for models with recombination hotspots) each consisting of 2000 independent 250 kb windows in a sample of 40 chromosomes where $\mu=1 \times 10^{-8}$ per bp per generation. Note, that when generating test datasets, we placed a Poisson number of mutations onto the genealogies in the usual fashion (Hudson 1983), rather than using the "fixed S approach" as we did for the

simulations used to estimate $\mathbf{p}$. We selected a subset of 20 SNPs (n_{snp}=20) with MAF $\geq 10\%$ from each window and constructed the observed *HCN* statistic for each dataset. We repeated this process 10 times for each dataset and used the average *HCN* statistic for inference.

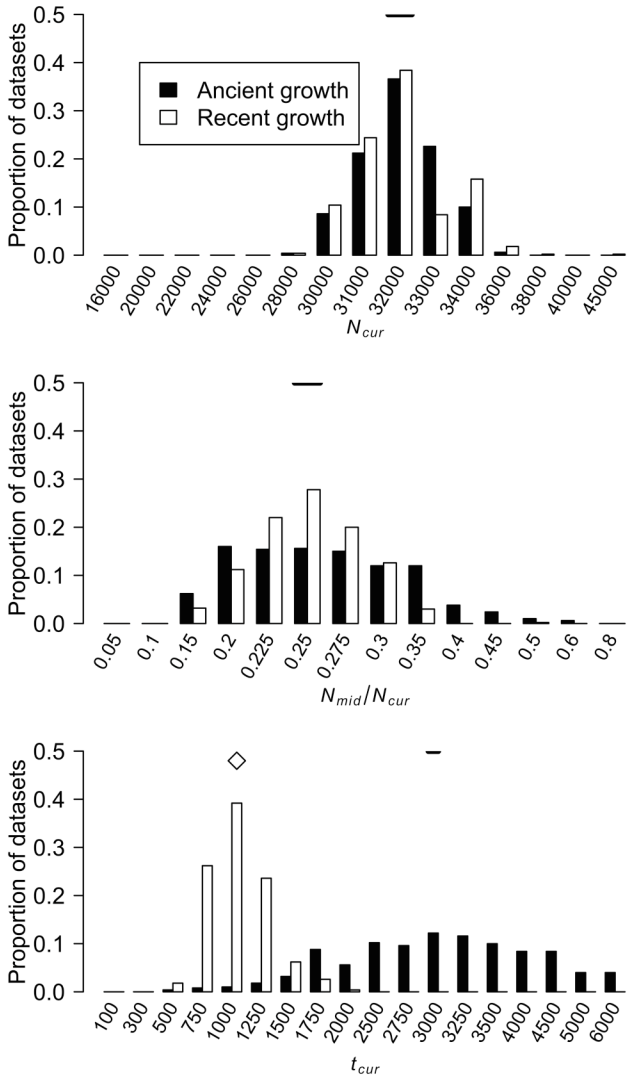


Figure 1.4: Distributions of MLEs of the three growth model parameters for simulated datasets under ancient growth and recent growth with uniform recombination (see Methods). The true value of each parameter is denoted by the horizontal line in each figure. Since t_{cur} differs between the two growth models, the true value of t_{cur} is denoted by a diamond for recent growth and a solid line for ancient growth.

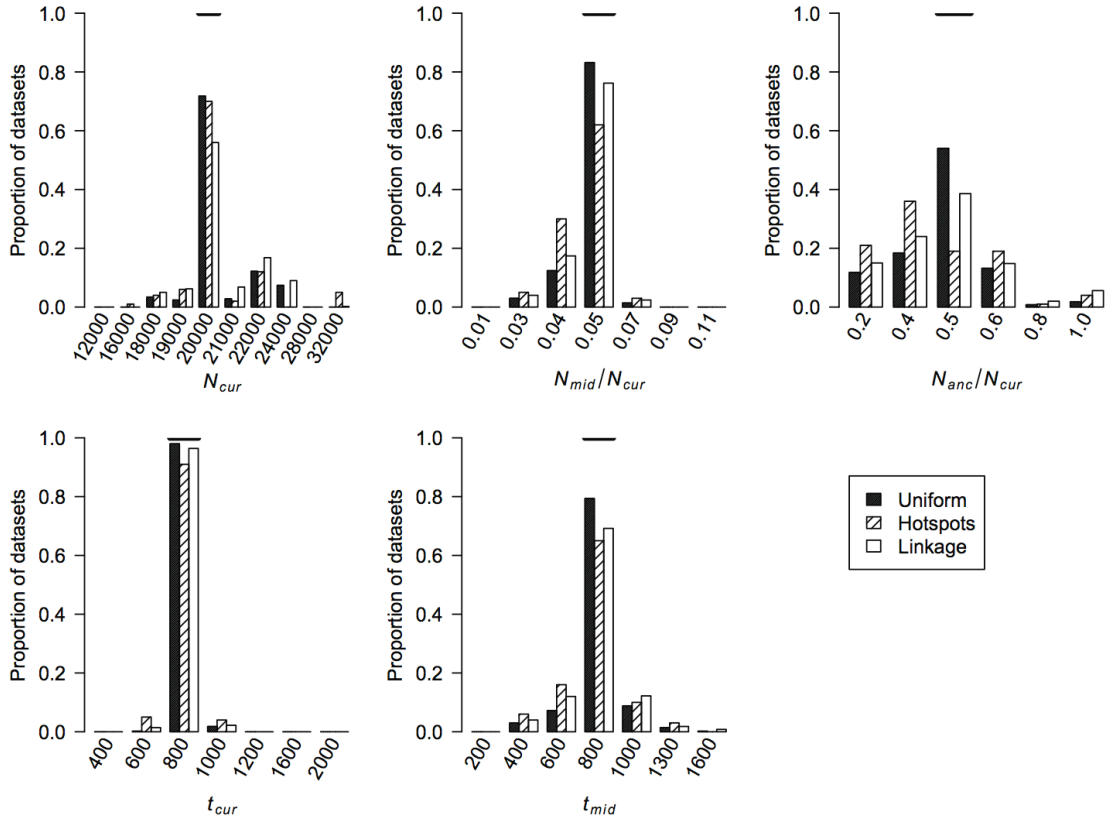


Figure 1.5: Distributions of MLEs of the five bottleneck model parameters for simulated datasets under uniform recombination, hotspots, and where some windows in the simulated datasets are linked to one another (see Methods). The true value of each parameter is denoted by the horizontal line in each figure.

For computational efficiency, we performed the coalescent simulations to estimate $\mathbf{p}$ over a grid of parameters (3780 and 20,580 parameter combinations for the growth and bottleneck models, respectively) once for each demographic and recombination model and stored the values to be used on subsequent datasets. The grid points used for each parameter are shown as the breaks in Figures 1.4-1.7. For each grid point we used 10^4 coalescent simulations to approximate the likelihood. Using a representative dataset, we then selected at least the top 10^3 grid points and ran an additional 10^5 replicates and for points near the MLE, we ran an additional 10^6 replicates. Due to computational constraints, these grids were not as dense as those used to estimate parameters in the Perlegen data.

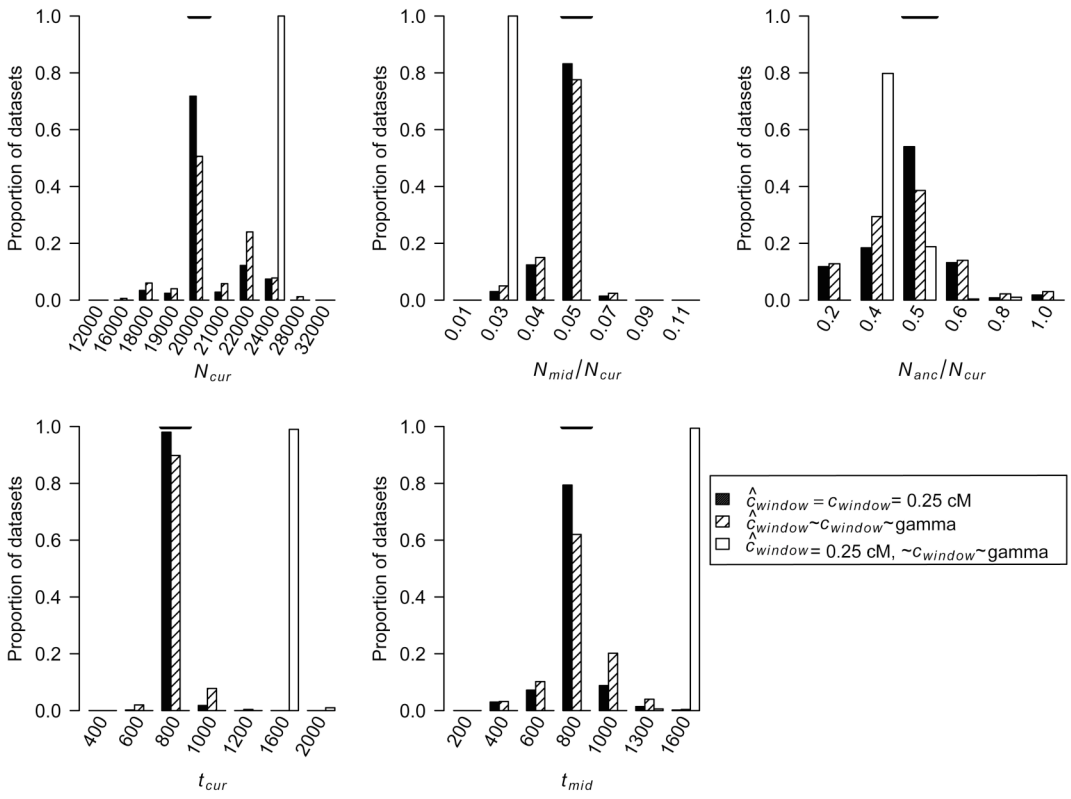


Figure 1.6: Distributions of MLEs of the five bottleneck model parameters for simulated datasets where there are errors in the genetic map.
$\hat{c}_{window} = 0.25$ cM, $c_{window} \sim gamma$ denotes the case where there are errors in the estimated genetic map that are ignored when performing the inference.
$\hat{c}_{window} \sim c_{window} \sim gamma$ denotes the case where we allow for errors in the genetic map when conducting the inference. The true value of each parameter is denoted by the horizontal line in each figure.

For all datasets and demographic models, the total amount of recombination within each window simulated was 0.25 cM (*i.e.* $c_{window} = 0.25$ cM). However, we also considered a model where there were 5 recombination hotspots present at random locations throughout each window. Each hotspot was 2 kb in size. The recombination rate (cM/bp) of each hotspot was drawn from a gamma distribution (shape=0.5, scale=2×10^{-6}). We then re-scaled the recombination rate of each hotspot such that 80% of the total amount of recombination in the window occurs within hotspots. We then assumed this model of recombination hotspots when inferring demographic

parameters. The test datasets were generated using the program msHOT (Hellenthal and Stephens 2007).

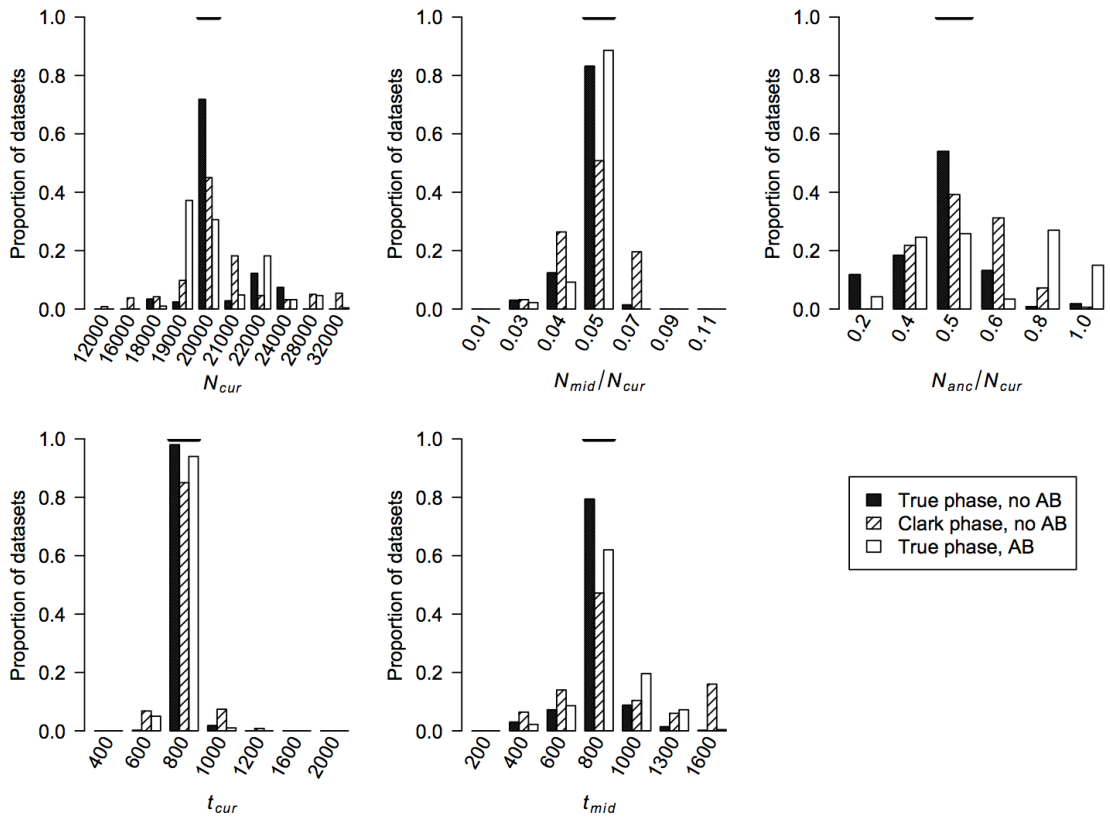


Figure 1.7: Distributions of MLEs of the five bottleneck model parameters for simulated datasets when phasing genotype data using Clark's phasing algorithm or there is SNP ascertainment bias (AB; see Methods). The true value of each parameter is denoted by the horizontal line in each figure.

Our method assumes that all windows in the genome are independent of each other. To assess the performance of our method when the windows are not independent, we simulated an additional 500 datasets using the same bottleneck model with $\hat{c}_{window} = c_{window} = 0.25$. Here the 2000 windows within each dataset were from 300 independent sets of 6-7 contiguous windows. The 2000 windows were treated as independent in the inference.

While for most of the simulations we assumed that there was no error in estimated recombination rates (*i.e.* $\hat{c}_{window} = c_{window}$), we also determined what effect

errors in the estimated genetic map had on our ability to accurately infer demographic parameters. Specifically, we simulated datasets under the same bottleneck model described above, but here instead of having $\hat{c}_{window} = c_{window} = 0.25$ cM, we drew c_{window} for each window from a gamma distribution (shape=10, scale=0.025). From this distribution, ~10% of windows will have c_{window} <0.155 and ~10% of windows will have c_{window} >0.355. We then inferred demographic parameters when incorrectly fixing $\hat{c}_{window} = 0.25$ cM. We also correctly incorporated errors into the genetic map by drawing $\hat{c}_{window}$ for each simulation replicate from the same gamma distribution used to generate the data.

Due to differences between the true *HCN* statistic and the *HCN* statistic constructed from phase-inferred haplotypes (Figure 1.8 and APPENDIX 1), it is important to incorporate the phasing process into the inference.

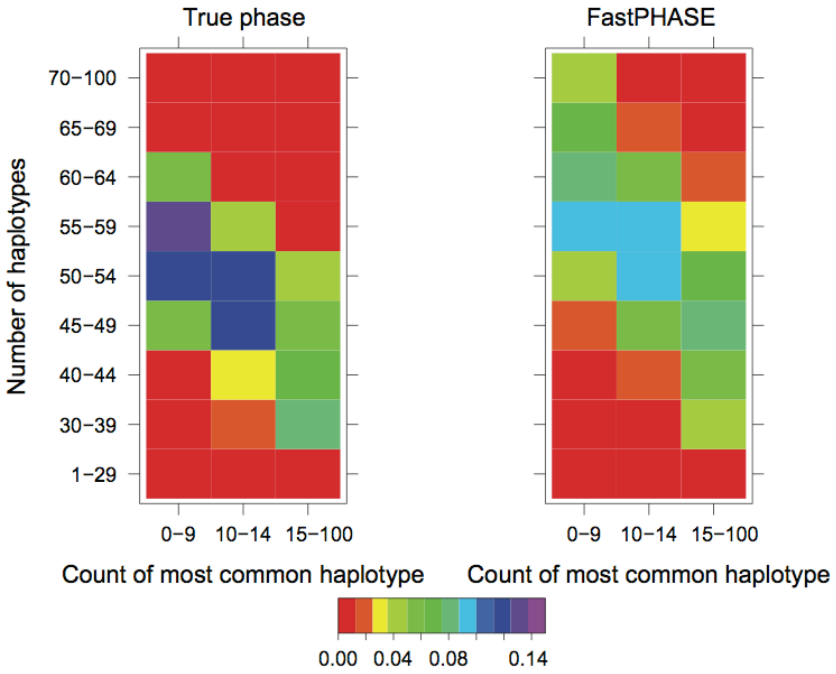
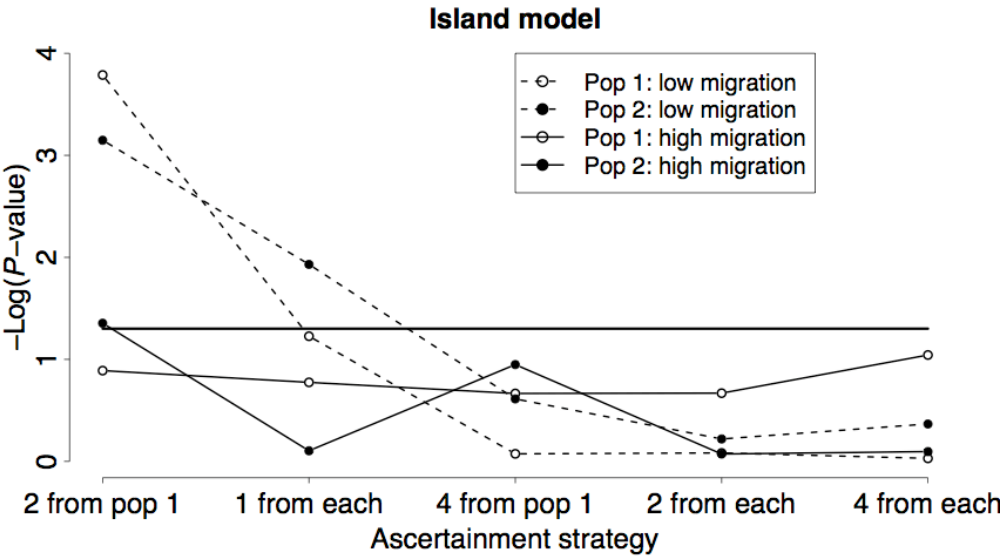


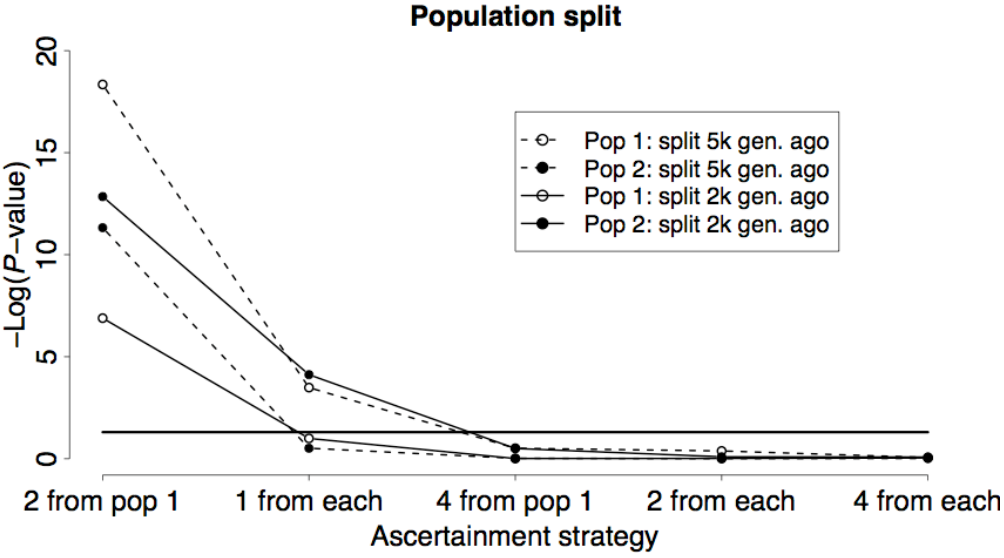
Figure 1.8: Effect of haplotype phase uncertainty on the *HCN* statistic. The *HCN* for a bottleneck model (see APPENDIX 1) when haplotype phase is known (left) and inferred using fastPHASE (right).

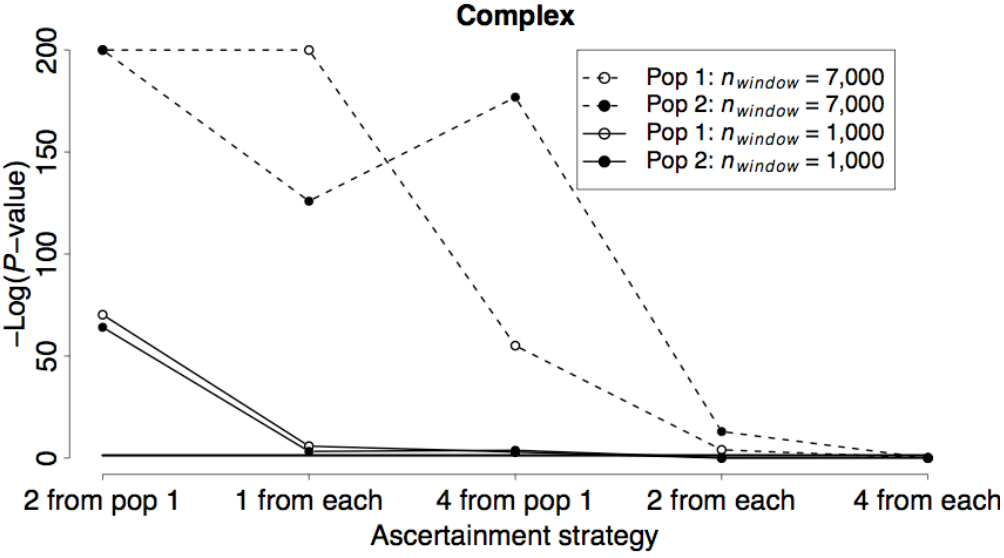
To do this, we suggest using the same phasing method that was used on the actual data to "phase" the simulated data used to estimate $\mathbf{p}$. Unfortunately, many phasing algorithms currently in use are computationally intensive and it would be nearly impossible to run these methods on the millions of coalescent simulation replicates used to find $\mathbf{p}$. For this reason, we examined the use of the computationally efficient parsimony phasing algorithm proposed by Clark (1990). If there are no individuals heterozygous at zero or one of the n_{snp} SNPs within a window or if there are genotypes that show no relation to known phased haplotypes, we arbitrarily assigned phase to a random individual and then use these two haplotypes to infer the rest. While this process may seem arbitrary, it can be done consistently both in the observed and simulated datasets. To make the method as computationally efficient as possible, we used only one ordering of the individuals. We assessed the performance of this approach by treating the simulated haplotypes in the test datasets as diploid genotypes and "phased" them using the parsimony method. For each simulation replicate to estimate $\mathbf{p}$ we also "phased" the simulated data using the same parsimony method.

Since we found that the *HCN* statistic constructed using SNPs that were discovered in a SNP discovery sample ≥ 8 chromosomes was very similar to the *HCN* statistic with complete SNP ascertainment (Figures 1.9-1.11; APPENDIX 1), we examined how ascertainment bias affected parameter estimates. Specifically, for the bottleneck model described above, we simulated a genotype sample of n=40 and as well as an additional SNP discovery sample of 6 chromosomes. Since the Perlegen SNPs were discovered using a multi-ethnic panel (Hinds *et al.* 2005), we included a SNP discovery sample of an additional 6 chromosomes from a second population (N_{cur}=10,000) that 5000 generations ago split from the population that underwent the bottleneck. To construct the *HCN* statistic from these simulated

Figure 1.9: Log_{10} *P*-value of the goodness-of-fit test comparing the *HCN* statistic under different SNP ascertainment schemes (shown on the *x*-axis) to that with complete ascertainment for three different demographic models (see APPENDIX 1). Here a sample size of 40 chromosomes from each population is used. The solid horizontal line in each figure denotes the 5% significance cutoff. *P*-values $<10^{-200}$ are set to 10^{-200}.


Island model
Pop 1: low migration
Pop 2: low migration
Pop 1: high migration
Pop 2: high migration
−Log(P−value)
0
1
2
3
4
2 from pop 1
1 from each
4 from pop 1
2 from each
4 from each
Ascertainment strategy


Population split
Pop 1: split 5k gen. ago
Pop 2: split 5k gen. ago
Pop 1: split 2k gen. ago
Pop 2: split 2k gen. ago
−Log(P−value)
0
5
10
15
20
2 from pop 1
1 from each
4 from pop 1
2 from each
4 from each
Ascertainment strategy


Complex
Pop 1: n_{window} = 7,000
Pop 2: n_{window} = 7,000
Pop 1: n_{window} = 1,000
Pop 2: n_{window} = 1,000
−Log(P−value)
0
50
100
150
200
2 from pop 1
1 from each
4 from pop 1
2 from each
4 from each
Ascertainment strategy

datasets, we only considered SNPs with MAF $\geq 10\%$ in the genotype sample that were variable in the 12 chromosome ascertainment panel. To infer parameters, we assumed there was no ascertainment bias (*i.e.* we used the same lookup tables for $\mathbf{p}$ that were described above that assumed complete SNP ascertainment).

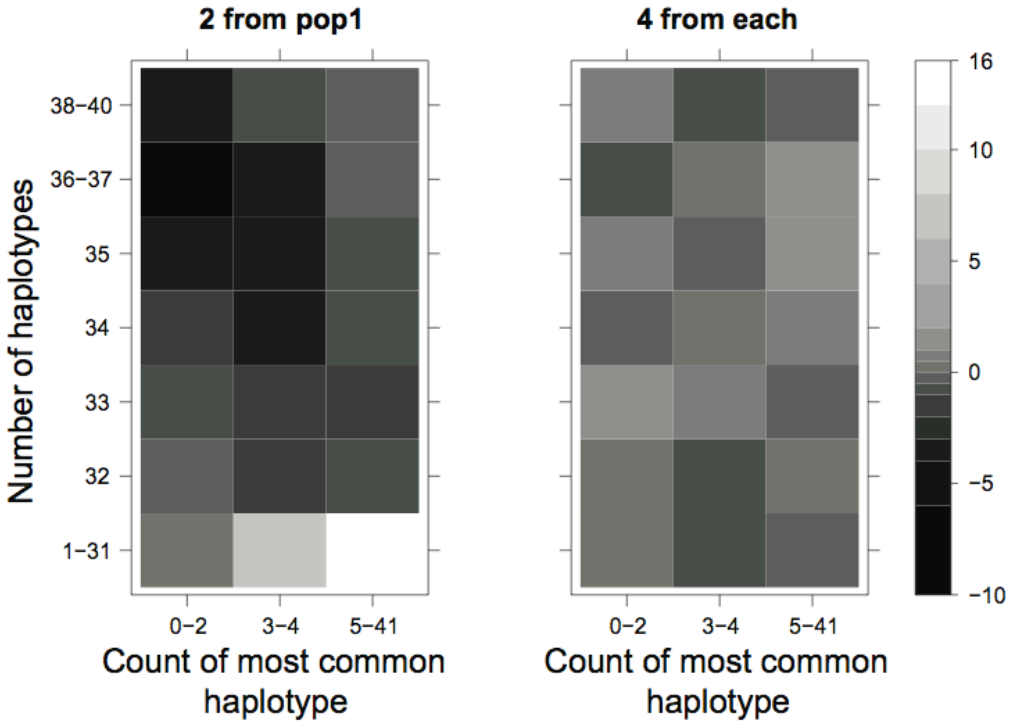


Figure 1.10: Plot of Pearson's residuals comparing the *HCN* statistic for two different ascertainment strategies to the expected *HCN* having complete SNP ascertainment for the bottlenecked population (population 1) in the complex demographic model. The two SNP ascertainment strategies compared are SNP ascertainment using 2 chromosomes from population 1 ("2 from pop 1") and ascertainment using 4 chromosomes from population 1 and 4 from population 2 ("4 from each"). Darker colors indicate a deficit of windows in the particular cell as compared complete ascertainment. Lighter colors indicate an excess of windows in the particular cell as compared complete ascertainment.

Analysis of Perlegen data

We applied our method to fit a bottleneck model to the CEU population genotyped by Perlegen Sciences (Hinds *et al.* 2005). We chose to use this population

since there is previous evidence of a bottleneck in this population (*e.g.* Marth *et al.* 2004; Voight *et al.* 2005), and all SNPs that were discovered by the Perelgen resequencing arrays were later genotyped in the CEU sample, without regard to LD status. We note that HapMap phase II specifically did not genotype SNPs that were in high LD in the Perlegen study, and this ascertainment criterion complicates the analysis of those data (see Discussion).

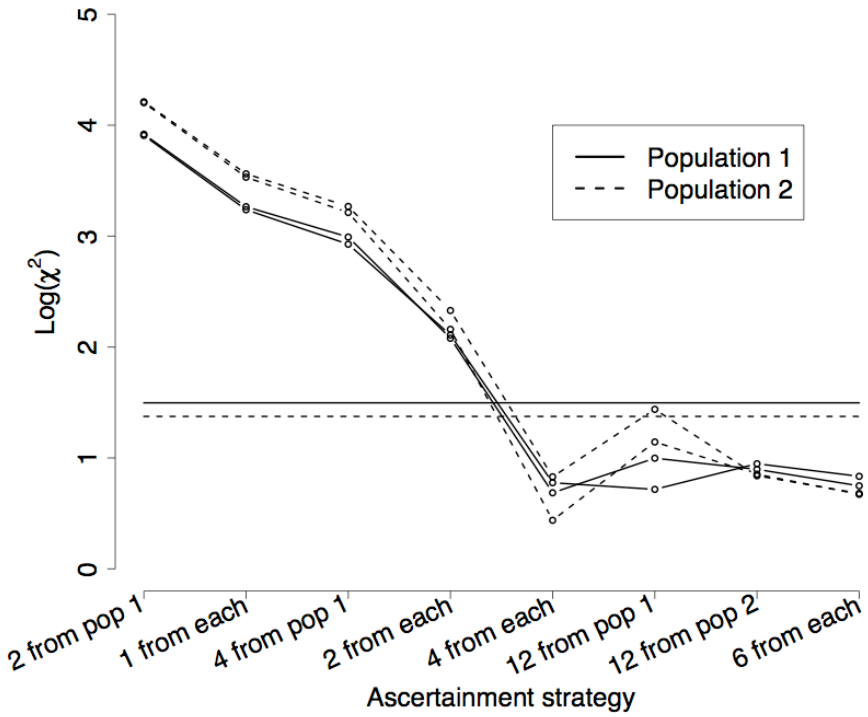


Figure 1.11: Log_{10} of the χ^2 statistic for the goodness of fit test comparing the *HCN* statistic under different SNP ascertainment schemes (shown on the *x*-axis) to that with complete ascertainment for the complex demographic model. Here a sample size of 120 chromosomes from each population is used. Note that the SNP discovery sample sizes used here differ from those in Figures 1.2 and 1.9. The horizontal lines denote the 5% significance cutoff for population 1 (solid) and population 2 (dashed). The two lines for each population are from two entirely independent replicates of the entire process (see APPENDIX 1) to assess stochastic variance.

We only considered autosomal (not X chromosome or mtDNA) SNPs with MAF $\geq 10\%$ in both the CEU and African American samples. Since our simulations of

ascertainment bias suggest that SNPs needed to have been discovered from discovery sample sizes >2 chromosomes, we only used those SNPs that were discovered in Perlegen's re-sequencing arrays of the multi-ethnic diversity panel (type "A" SNPs). There were 615,415 SNPs that fit both of these criteria. We used Clark's parsimony phasing algorithm to phase haplotypes in both the real data and in the simulation replicates to generate $\mathbf{p}$. For each population and in each window of the genome, we selected 10 random sub-sets of n_{snp} SNPs and constructed 10 different *HCN* statistics. We then used the average *HCN* statistic for inference.

We then set c_{window} = 0.25 cM and n_{snp} = 20. We used the LDhat genetic map (International HapMap Consortium 2007) to define windows since the deCODE map (based on pedigrees; Kong *et al.* 2002) does not have sufficient resolution for the scale of 0.25 cM (Myers *et al.* 2005). Since the quality of the genetic map used to divide the genome into windows can impact the inference, we drew $\hat{c}_{window}$ from a gamma distribution (scale=10, shape=0.025) to model errors in the genetic map. This distribution has a mean of 0.25 and variance of 0.00625. For the CEU data, n_{window}=8833.

We used a hotspot model similar to that of Schaffner *et al.* (2005; termed the "Schaffner hotspot model"). All hotspots had width of 2 kb. For each simulated window, hotspots occurred at random intervals drawn from a gamma distribution (shape=0.3, scale=8500/0.3) giving a mean spacing of 8500 bp (variance of ~2.41 x 10^8). Then the recombination rate (cM/2 kb) of each hotspot was drawn from another gamma distribution (shape = 0.3, scale =$c_{window}/0.3L$), where L is the physical size of the simulated window. In practice, L was drawn from the empirical distribution of physical distances for the n_{window} windows. We then re-scaled the recombination rate in the hotpots such that 88% of c_{window} occurs within hotspots. The amount of recombination occurring outside of hotspots is then equal to $0.12c_{window}$. Figures 1.12

and 1.13 show that this hotspot model matches the mean, standard deviation, and overall distribution (tabulated across all windows of the genome) of the observed inter-SNP genetic map distances quite well.

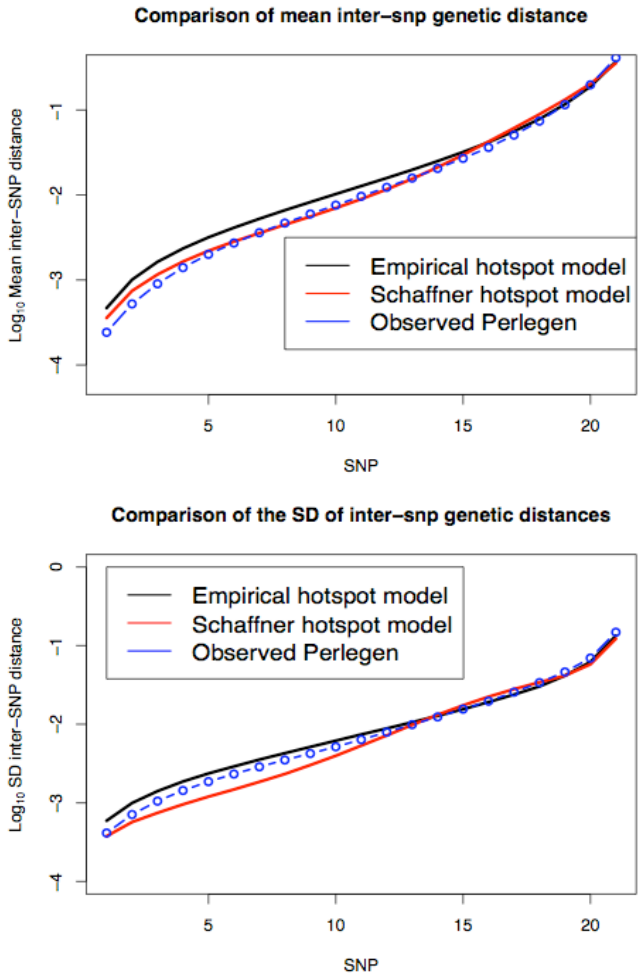


Figure 1.12: Comparison between the mean and standard deviation (SD) across all 8833 windows of the observed inter-SNP genetic distances (as defined by the LDhat genetic map) and the mean genetic distances simulated using the modified Schaffner hotspot model and the empirical hotspot model (see Methods). The left-most point in the top figure represents the mean of the smallest inter-SNP distance, averaged over all windows, the second point, the second smallest inter-SNP distance, and so on. The actual *HCN* statistic used for inference was averaged over 10 different *HCN* statistics, each of which was generated from a different random sub-set of SNPs from each window (see Methods). Here the observed and simulated inter-SNP genetic distances are based on selecting one random set of SNPs per window. The simulated inter-SNP genetic distances were determined assuming a constant population size, N=10,000, and re-scaling genetic distance for each window such that $\hat{c}_{window} = 0.25$ cM.

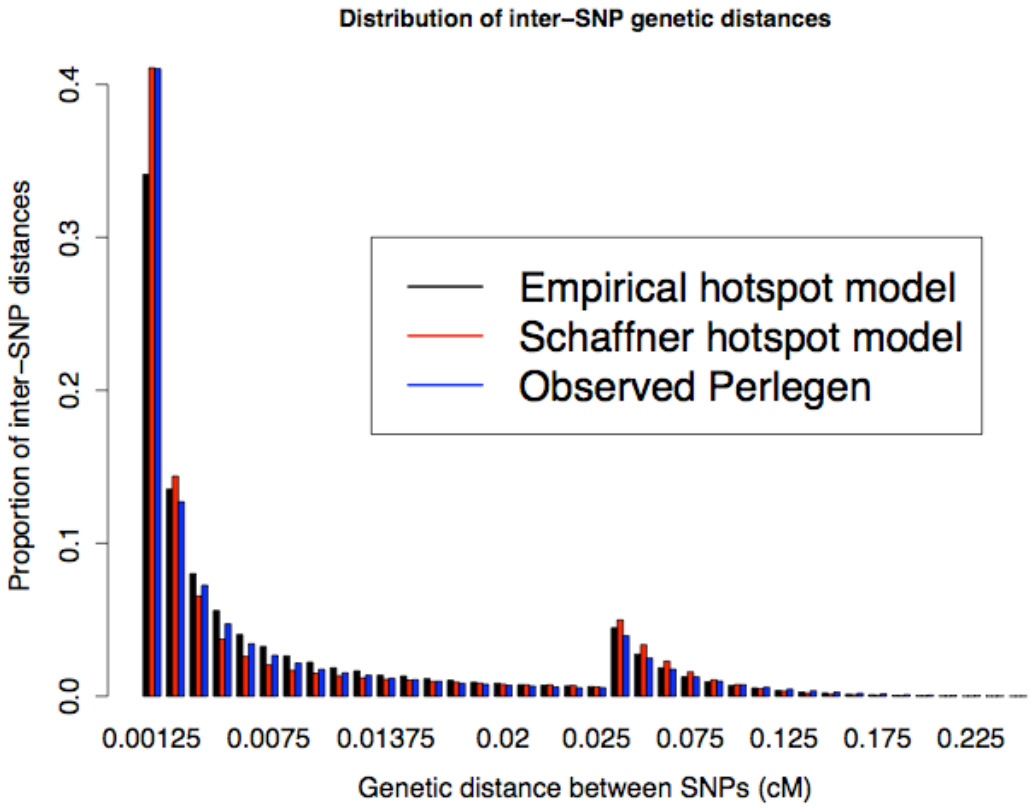


Figure 1.13: Comparison of the distribution of inter-SNP genetic distances in the Perlegen data (from the LDhat genetic map) with the Schaffner and empirical hotspot models (see Methods). The distribution is tabulated over all 8833 windows across the genome. The increased proportion in the bin after 0.025 cM is due to the change in scale of the bins. As noted in Figure 1.12, here the observed and simulated inter-SNP genetic distances are based on selecting one random set of SNPs per window. The simulated inter-SNP genetic distances were determined assuming a constant population size, N=10,000, and re-scaling genetic distance for each window such that $\hat{c}_{window} = 0.25$ cM.

In addition to the Schaffner hotspot model described above, we also directly used the estimated fine-scale LDhat genetic map (International HapMap Consortium 2007) as a guide to how recombination rates vary within windows (termed the "empirical hotspot" model). To do this, for each simulation replicate to estimate $\mathbf{p}$, we selected one of the 8833 windows at random and used the corresponding LDhat genetic map to delimit the relationship between genetic and physical distance for that replicate. We smoothed the map by only allowing the recombination rate to change at

points >500 bp and >0.0001 cM apart. We note that this hotspot model does not match the observed inter-SNP genetic distances as well as the Schaffner hotspot model does (Figures 1.12 and 1.13).

The grid to optimize the five-dimensional approximate likelihood function consisted of 85,536 points for the Schaffner hotspot model and 101,088 points for the empirical hotspot model. We used 12,500 simulation replicates for all points and 10^5 replicates for at least the top 4,000 points, and finally 10^6 replicates for at least the top 500 points. We found approximate 95% confidence intervals (CIs) for single parameters using asymptotic theory (*i.e.* the CI included points <1.92 log-likelihood units from the maximum), with linear interpolation of the profile likelihood curve to find points not directly simulated.

1.4 Results

Performance on simulated data

Figure 1.4 shows the distribution of the approximate maximum likelihood estimates (MLEs) of the three growth model parameters for simulated datasets under ancient growth (black bars) and recent growth (white bars). In both cases, the method is relatively unbiased for all three parameters. For ancient growth, N_{cur} is estimated most accurately, and t_{cur} the least. For recent growth, all three parameters are equally accurate, although for any given parameter, the MLE is the true value ~40% of the time. Notice that the variance in the distribution of MLEs for t_{cur} is much higher for ancient growth as compared to recent growth (making it the least precise as well as least accurate). Table 1.1 shows that for both growth scenarios in 100% of the datasets, the true parameter values were within the asymptotic 95% CIs (<3.9 log-likelihood units) around the MLE. Additionally, in >95% of the datasets, the single-parameter 95% CIs for all three parameters from the profile likelihood curves contained the true parameter values.

We also estimated the five parameters for a bottleneck model on simulated datasets. Figure 1.5 shows the distribution of the MLEs for the five bottleneck parameters. For the case of uniform recombination and $\hat{c}_{window} = c_{window} = 0.25$, the mode of the distribution MLEs for each parameter is at the true value of the parameter. The distribution of MLEs is tightest for N_{mid}/N_{cur} and t_{cur}, and broadest for N_{anc}/N_{cur}. This suggests that the recent bottleneck greatly alters haplotype patterns such that its timing and severity can be accurately estimated, but so much so that less information about the ancestral, pre-bottleneck population size (N_{anc}/N_{cur}) remains. Furthermore, the method appears to be relatively unbiased since it over- and under-estimates the true parameter value roughly equally. Table 1.1 shows that 99.8% of the time, the true parameter values within the asymptotic 95% CIs (within 5.5 log-likelihood units) around the MLE.

We next evaluated whether our method could accurately estimate demographic parameters in the presence of recombination hotspots (see Methods for the recombination hotspot model used). Figure 1.5 shows that when properly modeling recombination hotspots, we are able to accurately estimate the five bottleneck model parameters. Note that the distributions of the MLEs for all parameters have larger variances for the uniform recombination case. This pattern is due to the extra noise added by recombination hotspots. If a window of the genome has a low number of haplotypes and/or a high count for the most common haplotype, this could be due to demography (which is the only factor considered in the uniform recombination model) or due to SNPs falling in recombination coldspots. Consistent with this observation, Table 1.1 shows that a smaller proportion of the parameter space (99.65% vs. 99.87%) is >5.5 log-likelihood units from the MLE when there are recombination hotspots, as compared to uniform recombination. Notably, however, the method still appears to be

Table 1.1: Comparison of MLEs to the true parameter values for simulated datasets.

Recombination Model	Mean ll_{MLE}-ll_{Truth}[a]	Max ll_{MLE}-ll_{Truth}[b]	% MLEs = Truth[c]	Coverage of multi-D 95% CIs[d]	Coverage of 1D 95% CIs[e]	% points outside 95% CIs[f]
A. Ancient Growth						
$\hat{c}_{window} = c_{window} = 0.25$	0.631	3.47	3.0	100.0	99.87	94.79
B. Recent Growth						
$\hat{c}_{window} = c_{window} = 0.25$	0.437	3.09	22.6	100.0	99.93	98.84
C. Bottleneck						
$\hat{c}_{window} = c_{window} = 0.25$	0.363	6.31	47.4	99.8	99.52	99.87
Hotspots[f]	0.505	3.43	17.0	100.0	99.80	99.65
Linkage	0.732	8.30	29.8	99.4	98.48	99.87
$\hat{c}_{window} = 0.25$ $c_{window} \sim gamma$	136.685	179.07	0.0	0.0	8.36	99.98
$\hat{c}_{window} \sim c_{window} \sim gamma$	0.623	6.54	26.4	99.6	99.40	99.72
Clark's phasing algorithm[g]	0.779	4.90	19.8	98.0	99.96	99.43
Ascertainment bias	1.998	12.65	14.6	94.0	97.64	99.86

[a.] The average over all datasets of the log-likelihood at the MLEs minus the log-likelihood of the true demographic parameters.
[b.] The maximum distance between the log-likelihood at the MLEs and the log-likelihood of the true demographic parameters.
[c.] The proportion of datasets where the MLEs for all parameters were the true demographic parameters.
[d.] The proportion of datasets where the true parameters were <3.9 or <5.5 log-likelihood units of the MLEs, for the growth and bottleneck models, respectively.
[e.] The proportion of datasets where the true value of each parameter was <1.92 log-likelihood units from the MLEs using the profile log-likelihood curve, averaged over three or five parameters for the growth and bottleneck models, respectively.
[f.] The fraction of grid points (see Results) having a log-likelihood >3.9 or >5.5 log-likelihood units, for the growth and bottleneck models, respectively, from the MLEs.
[g.] Each window has 5 recombination hotspots, but for the whole window $\hat{c}_{window} = c_{window} = 0.25$ cM.
[h.] Haplotype phase was inferred in the test datasets and simulations to estimate **p** using Clark's phasing algorithm (see Methods).

unbiased, and for all 100 datasets, the true parameter values were <5.5 log-likelihood units from the MLE.

The simulations described above assumed that the 2000 windows in each dataset were independent. In practice, the windows may be contiguous along the genome, and thus are not independent. We examined the performance of our method on simulated datasets where some of the windows were linked. Figure 1.5 shows that for the distribution of the MLEs for certain parameters have greater variances than when the regions are unlinked, which is not surprising since 2000 linked windows contain less information than 2000 independent windows. As shown in Table 1.1, in all cases except when errors in the genetic map are ignored or there is SNP ascertainment bias (see below), the true parameter values for over 99% of the test datasets are within the asymptotic 95% CIs (<3.9 and <5.5 log-likelihood units from the MLE for growth and bottleneck models, respectively). This result suggests that the asymptotic CIs may actually be conservative, since the true parameter values are contained within the interval >95% of the time. When examining datasets where some of the regions are linked, we find that for 99.4% of the time, the true values are <5.5 log-likelihood units from the MLE (Table 1.1). Since in many cases, the 95% CIs for individual parameters based on the profile log-likelihood curve also appeared conservative (Table 1.1), we assessed their coverage in the datasets with linkage. For each of the five parameters, the true parameter value was <1.92 log-likelihood units from the MLE in at least 96.8% of the datasets. These results suggest that for the level of non-independence among windows, size of datasets, and parameter grid considered here, the asymptotic 95% CIs remain conservative.

The above simulations assumed that $\hat{c}_{window} = c_{window}$. In practice, $\hat{c}_{window}$ is estimated from a genetic map, either based on patterns of LD or pedigrees. We next evaluated the performance of our method when c_{window} is drawn from a gamma

distribution to mimic errors in the estimated genetic map. We first assumed that $\hat{c}_{window} = 0.25$ when running the inference (*i.e.* we ignored the errors in the genetic map). Figure 1.6 shows that our method performs poorly compared to the case where the genetic map is known with certainty. In particular, it overestimates t_{cur} and t_{mid}. Due to the fact that some windows in the simulated datasets will have low recombination rates, these windows will have very few haplotypes and a high frequency of the most common haplotype because c_{window} <0.25. Since we did not account for this in the inference, the method assumes that these low diversity windows were due to a stronger (or longer) bottleneck. Table 1.1 shows that the true parameter values are nowhere near the MLEs in this case. If, however, during the inference, $\hat{c}_{window}$ for each window is drawn from the same gamma distribution that generated the data, the method performs substantially better (Figure 1.6), though not quite as well as when $\hat{c}_{window} = c_{window} = 0.25$. Likewise, 99.6% of the time, the true parameter values were <5.5 log-likelihood units from the MLE. Note that, similar to what was seen for the case of recombination hotspots, on average, a smaller proportion of the parameter space (99.72% vs. 99.87%) is >5.5 log-likelihood units from the MLE when $\hat{c}_{window}$ and c_{window} follow a gamma distribution instead of being fixed at 0.25 cM.

To properly correct for errors introduced from inferring haplotype phase, we decided to phase the simulations used to estimate $\mathbf{p}$ using the same method as that used on the real data. We evaluated the performance of this strategy using Clark's phasing algorithm (Clark 1990) on 500 simulated datasets. Figure 1.7 shows the distribution of the MLEs for the five bottleneck parameters. This strategy works reasonably well and for each parameter the mode of the distribution of the MLE is at the true parameter value. Notice that the distributions of the MLEs for the datasets phased using Clark's phasing algorithm are broader than those when haplotype phase in known with certainty. Additionally, a smaller proportion of the parameter space is

excluded (>5.5 log-likelihood units from the MLE) when using Clark's phasing algorithm as compared to known phase data (99.43% vs. 99.87%; Table 1.1). This finding illustrates that, compared to having phase-known haplotypes, some information is lost when computationally inferring haplotypes. However the method appears reasonably unbiased, and in all simulated datasets, the true parameter values are <5.5 log-likelihood units from the MLE (Table 1.1).

To determine if we could accurately estimate bottleneck parameters in the presence of SNP ascertainment bias, we simulated datasets where the n_{snp}=20 SNPs for each window were picked from those SNPs with MAF $\geq 10\%$ in the genotype sample and were variable in the 12 chromosome SNP ascertainment sample. Figure 1.7 shows that for N_{mid}/N_{cur}, t_{cur}, and t_{mid}, our method performs very well even in the presence of SNP ascertainment bias. The distributions of the MLEs for N_{cur} and N_{anc}/N_{cur} are more variable than when there is no ascertainment bias, and the modes of their distributions are not at the true parameter values, suggesting that MLEs of these parameters are less reliable in the presence of ascertainment bias. The 95% CIs constructed from the profile likelihood curves remain conservative for N_{mid}/N_{cur}, t_{cur}, and t_{mid}, but for N_{cur}, the 95% CI is no longer conservative (*i.e.* the true value is within the 95% CI only 91.8% of the time). Additionally, the five dimensional 95% CI is also slightly anti-conservative (Table 1.1). For larger datasets (consisting of 10,000 independent regions) the CIs for N_{cur}, N_{anc}/N_{cur} and the 5 dimensional CI become even more anticonservative and the CI for t_{mid} becomes slightly anticonservative (not shown), likely due the fact that the ascertainment model is mis-specified, which has a stronger effect as the size of the dataset increases.

Inference of bottleneck parameters for the Perlegen CEU population

We fit the five-parameter bottleneck model to the Perlegen CEU population. Table 1.2 shows the MLEs and approximate 95% CIs for the five parameters when

Table 1.2: Inferred bottleneck parameters for the CEU dataset.

Recombination model		N_{cur}	N_{mid}/N_{cur}	N_{anc}/N_{cur}	t_{cur}	t_{mid}
Schaffner[a]	MLE	10,000	0.055	0.8	1100	400
	95% CI	9315-10,665	0.052-0.064	0.70-?[c]	1017-1140	360-475
Empirical[b]	MLE	10,000	0.05	0.8	1200	300
	95% CI	9440-11,454	0.042-0.066	0.66-0.97	1086-1437	260-552

MLEs and approximate 95% CIs from the profile likelihood curves for the 5 parameter bottleneck model.

[a.] Hotspot model is similar to that described by Schaffner *et al.* (2005); See Methods.

[b.] Hotspot model is based directly on the LDhat genetic map.

[c.] We did not estimate an upper boundary on the CI since the profile likelihood surface for N_{anc}/N_{cur} is relatively flat from 0.8 to 1.0 and we did not consider points >1.0.

using both the Schaffner model of recombination hotspots as well as the empirical hotspot model based on the LDhat genetic map. Figure 1.14 shows that for both hotspot models, the *HCN*s generated using the MLE parameter estimates match the observed CEU *HCN* quite well. The current population size is estimated to be ~10,000 when using both recombination models. There was a severe population size reduction (~4.2-6.6% of the current size) lasting 260-552 generations (see Figure 1.15 for the 2-dimentional profile likelihood surface). The bottleneck began approximately 1017-1437 generations ago.

Based on the two-dimensional profile likelihood surface (Figure 1.16), we estimate that the bottleneck began (t_{cur}+t_{mid}) 1500 generations ago---37,500 years, assuming 25 years/ generation when using either hotspot model. Notably, the oldest start time within the asymptotic 95% CI (3 log-likelihood units, for 2 df) is 1600 generations, or 40,000 years for the Schaffner hotspot model, and 1900 generations (47,500 years) for the empirical hotspot model.

One potential concern with using SNPs from the Perlegen SNP discovery project is that it contains a lot of missing data, resulting in some fraction of SNPs having been discovered in a smaller sample. To determine what affect this had on the inference of the bottleneck parameters, we examined the depth of the SNP discovery panel for the 615,415 SNPs used in constructing the *HCN* statistic used for demographic inference in the Perlegen CEU sample. We found that 11.8% of these SNPs were discovered by comparing fewer than 8 chromosomes. We removed these SNPs and re-computed the *HCN* statistic from the Perlegen CEU data (now with 8174 windows) and re-estimated the bottleneck parameters using the Schaffner recombination hotspot model. We found identical MLEs for the parameters as in our original analysis.

Figure 1.14: Observed *HCN* statistic for the Perlegen CEU sample and the *HCN* statistics for the best-fitting demographic models based on the Schaffner hotspot model and the empirical hotspot model. Windows based on genetic distance were defined using the LDHat genetic map (see Methods). See Table 1.2 for the parameter values generating the best-fitting *HCN* statistics. Note, the bins shown in the figure were the same ones used when inferring parameters.

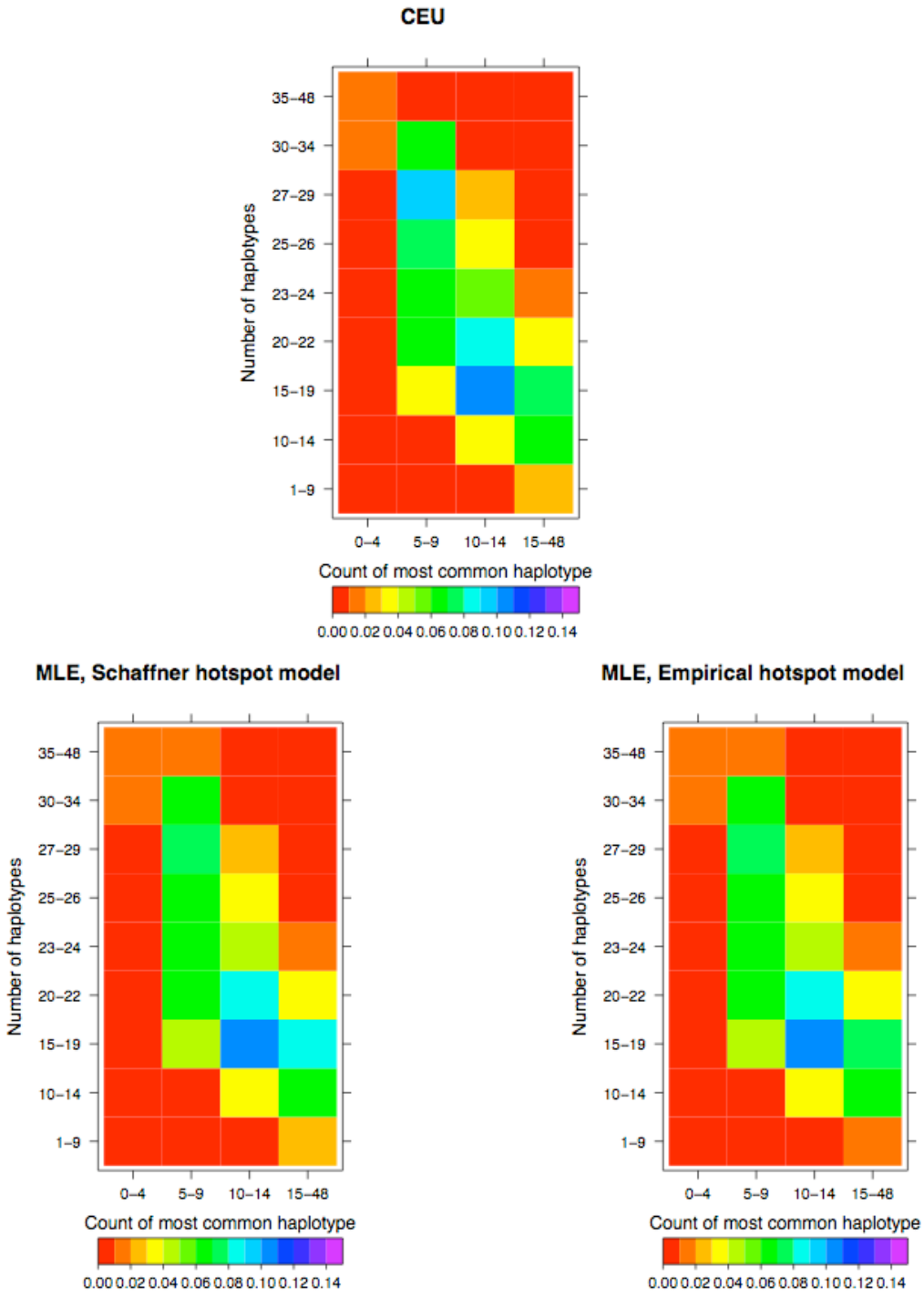
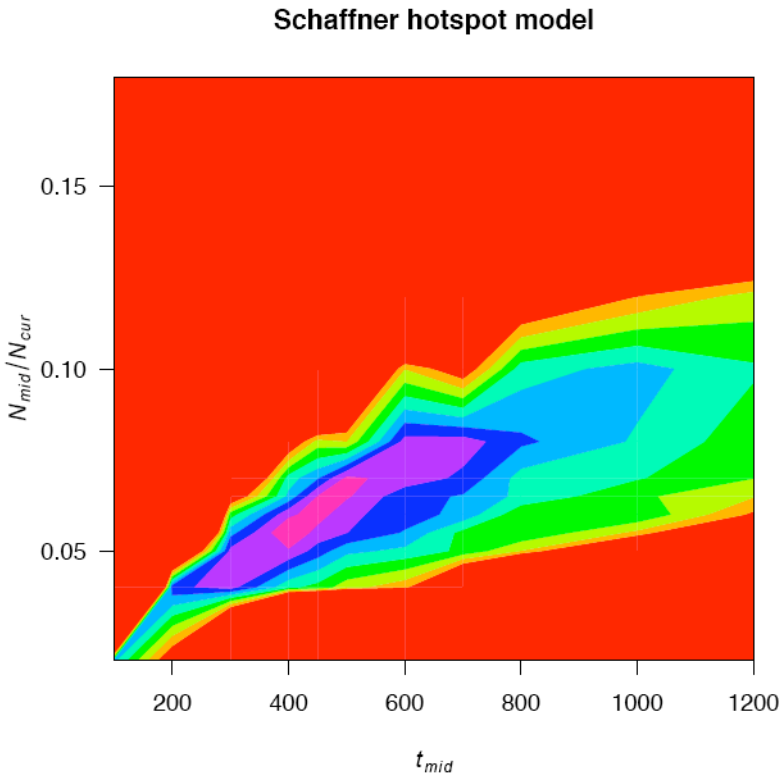

CEU
Number of haplotypes
35–48
30–34
27–29
25–26
23–24
20–22
15–19
10–14
1–9
0–4
5–9
10–14
15–48
Count of most common haplotype
0.00 0.02 0.04 0.06 0.08 0.10 0.12 0.14
MLE, Schaffner hotspot model
Number of haplotypes
35–48
30–34
27–29
25–26
23–24
20–22
15–19
10–14
1–9
0–4
5–9
10–14
15–48
Count of most common haplotype
0.00 0.02 0.04 0.06 0.08 0.10 0.12 0.14
MLE, Empirical hotspot model
Number of haplotypes
35–48
30–34
27–29
25–26
23–24
20–22
15–19
10–14
1–9
0–4
5–9
10–14
15–48
Count of most common haplotype
0.00 0.02 0.04 0.06 0.08 0.10 0.12 0.14

Figure 1.15: 2D-profile likelihood plot for t_{mid} vs. N_{mid}/N_{cur} for the Perlegen CEU data inferred using the Schaffner hotspot model and empirical hotspot model. Contours are every 3-log-likelihood units. The inner pink contour denotes the region of points where the log-likelihood is < 3-log likelihood units from the MLE.


Schaffner hotspot model
0.15
0.10
0.05
N_{mid}/N_{cur}
200
400
600
800
1000
1200
t_{mid}

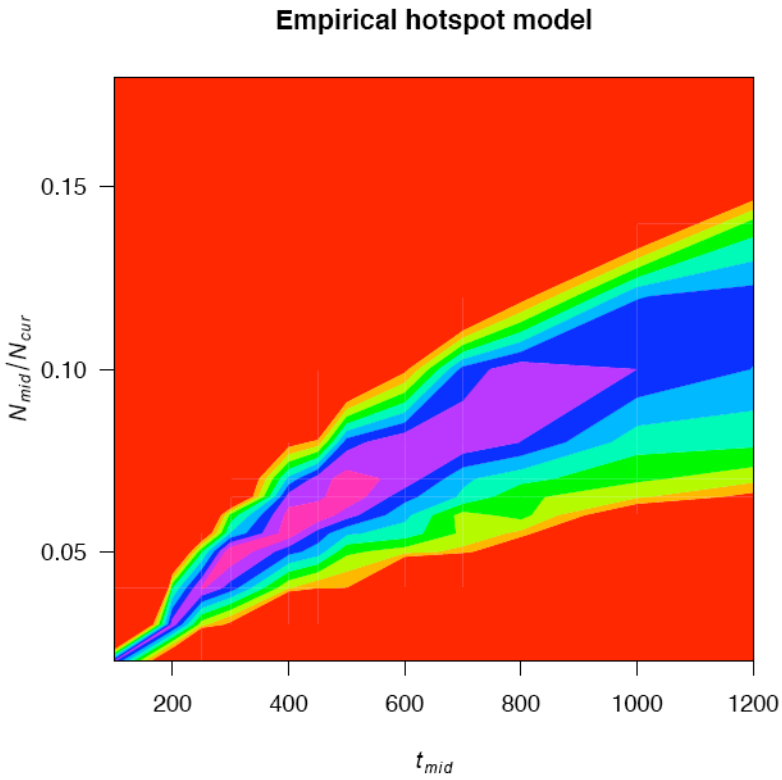
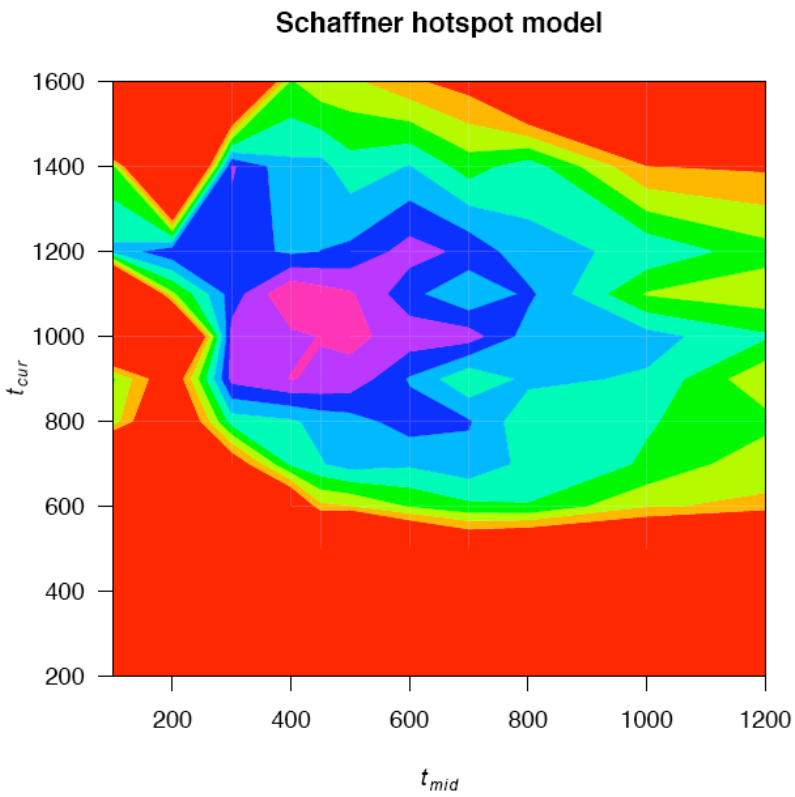
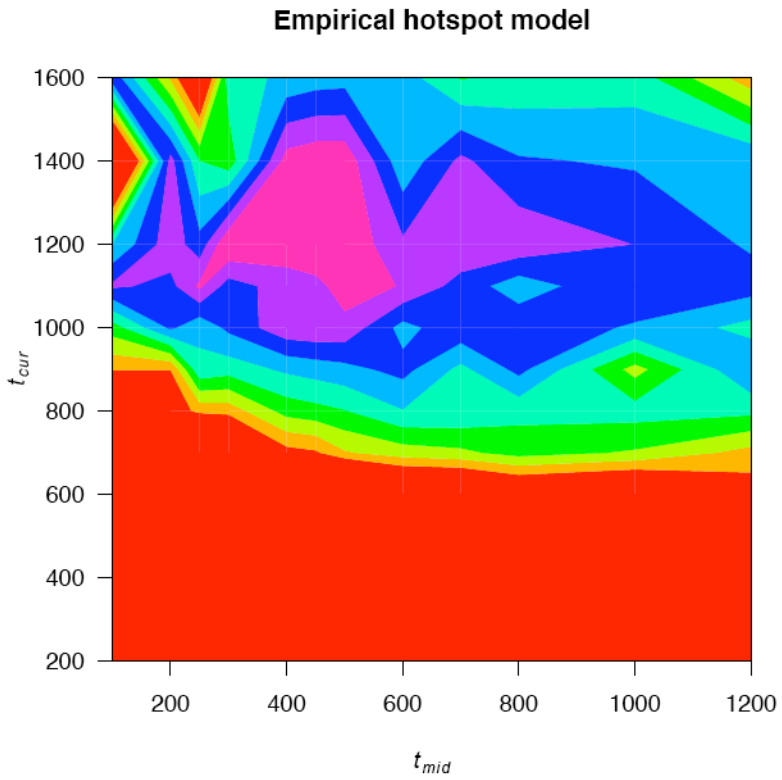

Empirical hotspot model
0.15
0.10
0.05
N_{mid}/N_{cur}
200
400
600
800
1000
1200
t_{mid}

Figure 1.16: 2D-profile likelihood plot for t_{cur} vs. t_{mid} for the Perlegen CEU data inferred using the Schaffner hotspot model and empirical hotspot model. Contours are every 3-log-likelihood units. The inner pink contour denotes the region of points where the log-likelihood is < 3-log likelihood units from the MLE. Note the jaggedness of the contours is due to the relatively course grid used to estimate parameters combined with Monte Carlo error.


Schaffner hotspot model
1600
1400
1200
1000
800
600
400
200
t_{cur}
200
400
600
800
1000
1200
t_{mid}


Empirical hotspot model
1600
1400
1200
1000
800
600
400
200
t_{cur}
200
400
600
800
1000
1200
t_{mid}

1.5 Discussion

We have proposed a flexible approximate likelihood method to estimate demographic parameters using haplotype summary statistics. We have shown that accurate estimates of demographic parameters can be made using genome-wide SNP datasets of practical size. To the best of our knowledge, this is one of the first studies to estimate parameters in a demographic model in a likelihood framework using haplotype patterns from genome-wide SNP genotype data. Furthermore, we have addressed many complications that arise in the analysis of genome-wide data, such as recombination rate heterogeneity, errors in the estimated genetic map, haplotype phase uncertainty, and ascertainment bias. Provided that good genetic maps are available, our method could be applied to SNP data from other species, such as dogs and cattle, to estimate domestication bottleneck parameters.

One of the major disadvantages of our approach is its dependence on accurately modeling the distribution of recombination rates across the genome. Our simulations have shown that errors in the genetic map can cause poor performance. Therefore, we suggest only applying our method when there is an accurate genetic map for the species in question. We suggest incorporating a distribution on c_{window} to allow for errors in the estimated genetic map. However, as the quality and resolution of genetic maps continue to improve, the utility and accuracy of our method will also continue to increase. For species where recombination rates vary at the fine scale, it is crucial to incorporate some model of recombination hotspots. Here for the Perlegen CEU data, we have implemented a parametric model as well as empirical estimates based on LD patterns. A similar influence of the assumed recombination rate on the demographic parameter estimates was noted in Thornton and Andolfatto (2006), who used the variance in the number of haplotypes across windows as one of their summary statistics. We recommend, as done in Thornton and Andolfatto (2006), using

different recombination models and then comparing the final results to assess how dependent the estimates are on the assumed recombination model. While the dependence of our method on accurate estimates of the recombination rate is not ideal, we point out that many previous methods in molecular evolution and population genetics are dependent on accurate estimates of the mutation rate, and will be biased if erroneous estimates are used.

We also assume that the genetic map remains constant over time and is the same across populations. Recombination hotspots do not appear in the same locations in chimps and humans despite a high level of sequence identity (Ptak *et al.* 2005; Winckler *et al.* 2005). It has therefore been speculated that recombination hotspots are not permanent features of the genome, and evolve on a time-scale of at least tens of thousands of years (Jeffreys *et al.* 2005). However, it appears that the time scale over which many hotspots evolve is older than 100,000 years, and because this is long enough to alter patterns of LD, temporal changes in hotspots on this timescale will not have such a severe impact on our method. Hotspots that evolve over shorter time scales, or are population specific may have a larger effect on our method. This effect is hard to quantify since the prevalence of rapidly evolving or population specific hotspots, other than the existence of a few examples (Jeffreys *et al.* 2005; Clark *et al.* 2007), remains largely unknown. Encouragingly, a recent paper (Hellenthal *et al.* 2006) found that genotype-specific recombination events would not substantially affect LD patterns, boding well for our method.

We have found that the *HCN* statistic constructed from computationally phase-inferred data differs from the true *HCN*. Simply treating the phase inferred haplotypes as the true haplotypes will likely give biased parameter estimates. Thus, when analyzing data from unrelated individuals, it is important to consider errors induced in the phasing process. We suggest doing this by inferring phase for the coalescent

simulations used to estimate *HCN*. Our simulations suggest that using Clark's phasing algorithm works well for this purpose. However, some information is lost by this procedure (Figure 1.7; Table 1.1), and we therefore recommend, where available, using data from trios, where haplotype phase can be inferred with great accuracy, to maximize the information in the data

It has been suggested (Conrad *et al.* 2006) that haplotype statistics may be less susceptible to SNP ascertainment bias than statistics based upon SNP frequencies. Here, we have extensively investigated whether this holds true for the *HCN* and H_{pair} statistics under a variety of demographic and ascertainment conditions. Encouragingly, we found that the *HCN* statistic is reasonably robust to SNP ascertainment bias provided that the SNP discovery sample is sufficiently deep. The reason for this is that we focus on subsets of common SNPs, rather than on rare SNPs. However, the H_{pair} statistic was very susceptible to ascertainment bias (Figure 1.2 and APPENDIX 1), suggesting that all haplotype statistics are not equally affected by ascertainment bias, and it will be necessary to explicitly evaluate, as we have done here, whether ascertainment bias affects a particular haplotype statistic.

The sizes of the SNP discovery and genotype samples play an important role in determining the effect of ascertainment bias on the *HCN* statistic. Interestingly, generating the *HCN* from SNPs ascertained uniformly using two chromosomes of known ethnicity (as done by Keinan *et al.* 2007) would result in a very different *HCN* statistic than that expected without ascertainment bias (Figures 1.9-1.11), which would result in biased inference. Though not considered here, it is in principle possible to estimate the expected *HCN* statistic conditional on this particular ascertainment strategy, and application of such an estimate would reduce this bias.

We have found that SNP discovery sample sizes of at least 12 total chromosomes should be sufficient to result in the *HCN* statistic from ascertained SNPs

to match the expected *HCN* when considering genotype samples of 40 or 120 chromosomes. Furthermore, we have shown that our method can reliably infer several bottleneck parameters when SNPs were ascertained in the manner. As the SNP discovery sample size increases, performance of our method would continue to improve and become closer to that for the case of no ascertainment bias, since a larger SNP discovery sample will capture more of the SNPs in the genotype sample. These results are especially encouraging since Perlengen's SNP discovery effort used 20-50 chromosomes where ~12 chromosomes were of African American ancestry and ~12 chromosomes were of European ancestry, with the remainder being Mexican American, Asian American, and Native American (Collins *et al.* 1998; Hinds *et al.* 2005).

Furthermore, we have found that the size of the SNP discovery sample is more important than whether or not the SNP ascertainment had been done in a particular population (see Figure 1.11). This suggests that SNPs that were ascertained in the Perlegen resequencing survey could be used to estimate demographic parameters in other populations not represented in the SNP discovery panel. It is worth noting that the two populations in our simulation study for Figure 1.11 split 5000 generations ago (125,000 years, assuming 25 years/generation) with no subsequent migration, and are thus more differentiated than many actual non-African populations that could be studied empirically.

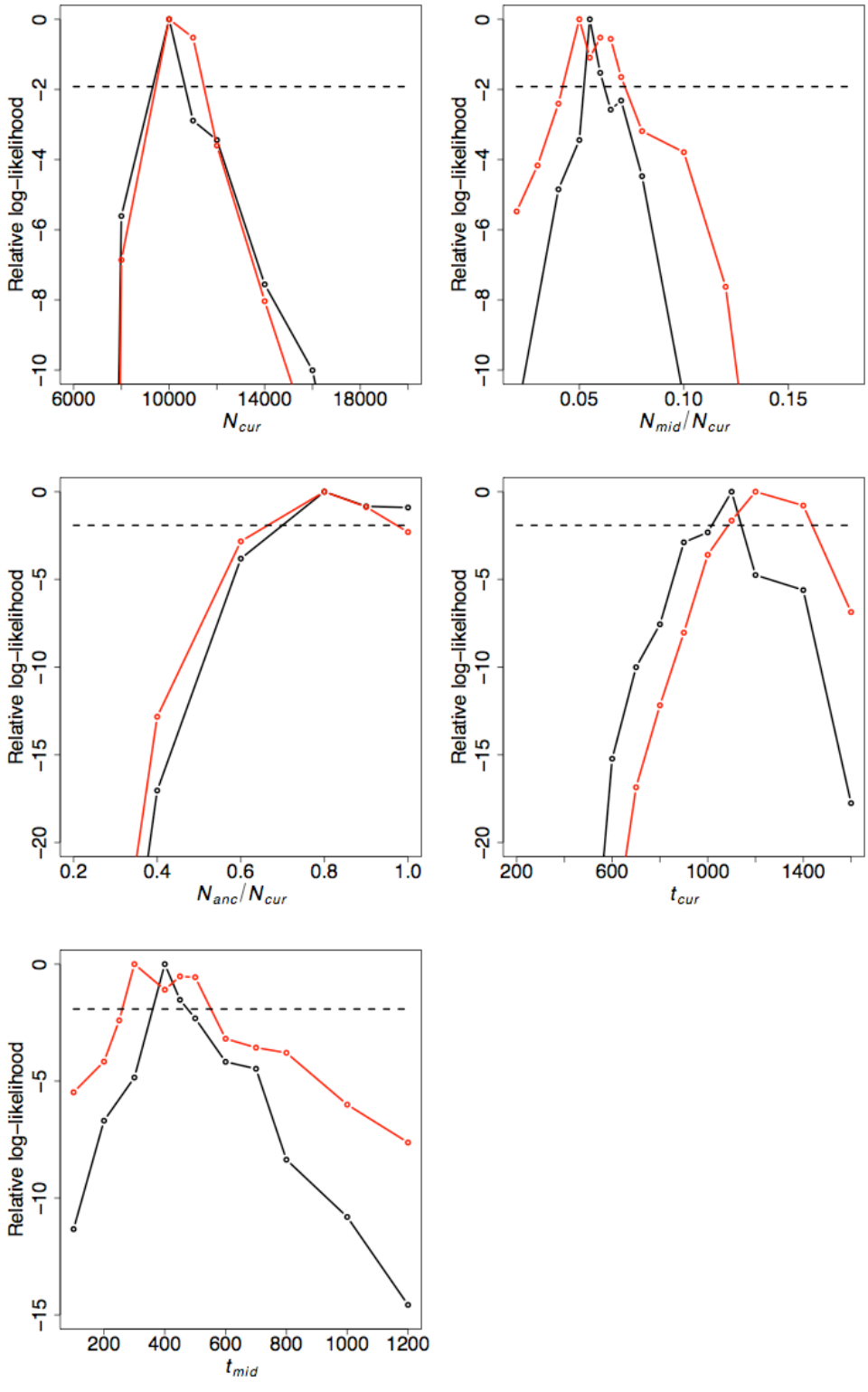
It is important to note that the type of ascertainment bias studied here is due to the preferential genotyping of common SNPs. In all the analyses presented here, we assume that the genotyped SNPs were selected without regard to physical or genetic distance or LD patterns. Such an assumption is reasonable for the analyses of the Perlegen data presented here since Perlegen attempted to genotype all of the SNPs found in their SNP discovery process (Hinds *et al.* 2005). The assumption is not valid,

however, for many of the "SNP chips" that preferentially selected SNPs based upon physical distance (in the case of Affymetrix 500k) or LD patterns (in the case of the Illumina platform; Eberle *et al.* 2007). Since the SNPs on these platforms are not a random subset of the total variation, using our method on such data will likely give misleading results. In principle, it should be possible to modify our method to model the SNP selection process in the inference, which would allow our method to be applied to the large-scale SNP genotype datasets that have been collected, such as the HGDP dataset (Jakobsson *et al.* 2008; Li *et al.* 2008).

For the analysis of the CEU data, we used two different models of recombination rate variation. Overall, the results using both models are qualitatively similar, suggesting that our method is somewhat robust to minor mis-specification of the recombination hotspot model. We also find that the single-parameter 95% CIs from the profile likelihood overlap for all five parameters estimated (Table 1.2; Figure 1.17). Nevertheless, the five-dimensional 95% CIs do differ between the two recombination models, mainly due to the fact that t_{cur} is greater under the empirical hotspot model than the Schaffner hotspot model.

The time we inferred that the CEU bottleneck began (~37, 500 years) ago is too recent to coincide with the accepted dates for the Out-of-Africa bottleneck, which is believed to have occurred 40,000-80,000 years ago (Reed and Tishkoff 2006). Thus, our estimate may coincide with an additional bottleneck associated with the founding of Europe, which likely took place 30,000-40,000 years ago (Barbujani and Goldstein 2004). Alternatively, our estimated start time of the bottleneck may represent an average time over several bottlenecks, including the Out-of-Africa bottleneck and a more recent bottleneck, perhaps associated with the Last Glacial

Figure 1.17: Likelihood profiles for the five CEU bottleneck parameters inferred using the Schaffner hotspot model (black) and the empirical hotspot model (red). The dashed line denotes the approximate 95% confidence interval.


Relative log−likelihood
0
−2
−4
−6
−8
−10
6000
10000
14000
18000
N_{cur}
0.05
0.10
0.15
N_{mid}/N_{cur}
−5
−10
−15
−20
0.2
0.4
0.6
0.8
1.0
N_{anc}/N_{cur}
200
600
1000
1400
t_{cur}
200
400
600
800
1000
1200
t_{mid}

Maximum which began ~18,000 years ago (Barbujani and Goldstein 2004). Further work considering multiple European populations and multiple bottlenecks may help resolve this question.

How do our estimates of the bottleneck parameters for the CEU match with published estimates? Voight *et al.* (2005) may not be directly comparable to our study since their analysis considered a Southern European sample and ours used individuals with Northwestern European ancestry, and differences in haplotype diversity between these two regions have been noted (Lao *et al.* 2008). Nevertheless, our estimates of t_{mid} and N_{mid}/N_{cur} fall within their confidence regions. Their bottleneck start times (40,000 years) and current and ancestral population size estimates (~10,000) also agree with ours. Our estimates of the time the bottleneck ended is also consistent with that found using the decay of pairwise LD. Reich *et al.* (2001) found evidence for a bottleneck 800-1600 generations ago (our MLE is 1500 generations). We find evidence for a more severe bottleneck than previously estimated (Adams and Hudson 2004; Marth *et al.* 2004; Keinan *et al.* 2007) which could reflect the importance of considering LD-based information in the inference since these studies are all based on the frequency spectrum, and the other study that considers summary of LD (Voight *et al.* 2005), cannot reject such a severe bottleneck. Alternatively, there could be some other important factors in European population history not captured by these simple bottleneck models which may affect the frequency spectrum and LD patterns differently. Finally, we cannot exclude the possibility that we have overestimated the bottleneck intensity due to greater heterogeneity in the recombination rate than what was included in our hotspot models. In short, improved confidence in the fine-scale genetic map will allow definitive ability to discriminate between these alternative scenarios.

While we have shown that haplotype statistics can be used to estimate demographic parameters from SNP genotype data, and it has been shown that the site frequency spectrum (SFS) will give misleading results when applied to genotype data without a correction for ascertainment bias (Nielsen *et al.* 2004; Clark *et al.* 2005), one important question is whether haplotype summary statistics will provide additional information that is important for inference when full genome-wide resequencing data are available and it is possible estimate the SFS accurately? We examined whether the *HCN* statistic discriminates between two different demographic models that have similar SFSs. We focused on a demographic model that included ancestral population structure since previous studies found that ancestral structure can result in an excess of long-range LD (Wall 2000b; Plagnol and Wall 2006). Specifically, we found that for certain subsets of the parameter space (the ms command lines giving the parameters used to generate Figure 1.18 are given in APPENDIX 1) population growth with ancestral structure can create a similar SFS to that expected under population growth without ancestral structure. Close inspection of Figure 1.18 reveals a very slight uptick in the proportion of high frequency derived SNPs in the population growth with structure SFS as compared to the growth without structure SFS, which is the expected signal of ancestral population structure. However, this effect is very subtle and in practice may be attributed to mis-identification of the derived allele, rather than ancestral population structure (Hernandez *et al.* 2007a). Note that while the magnitude of growth in the structure and panmictic cases are different, the growth with structure case still has an excess of low frequency SNPs (Figure 1.18) which would often be interpreted as evidence for population growth. The insert within Figure 1.18 shows the count of the most common haplotype versus the number of haplotypes for 10,000 windows simulated under the two demographic models described above.

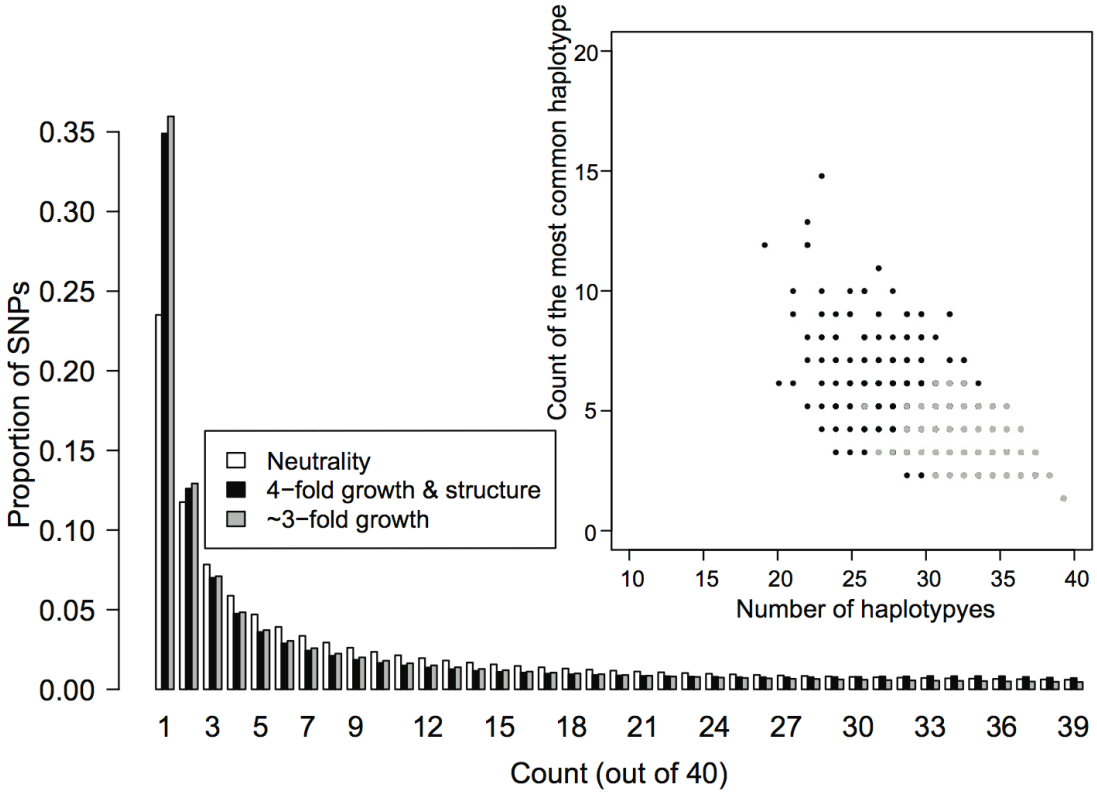


Figure 1.18: Comparison of the expected SFS for ancestral population structure combined with population growth to that expected with just population growth. The insert shows the frequency of the most common haplotype versus the number of haplotypes per simulated window for the same two demographic models. Note that the SFS for the two models appear similar, but that there is an excess of windows where the most common haplotype is at high frequency in the population structure combined with growth demographic model (see Discussion).

Note the growth with structure model has an excess of windows where the most common haplotype is at higher frequency and an excess of windows with a fewer number of haplotypes compared to the pure growth model. Thus, this is a case where two demographic models that cannot be readily differentiated on the basis of the SFS can be distinguished easily using haplotype patterns. The reason for this is as follows: population growth results in an excess of low-frequency SNPs, and for the parameters used here, population structure results in an excess of both low-frequency and high-frequency derived alleles. The resulting SFSs in Figure 1.18 have been affected by

both these forces, but the excess of low-frequency SNPs is the predominant feature. Since ancestral population structure results in some genealogies having longer internal branches, any mutations occurring on these branches will be in LD with each other, leading to fewer distinct haplotypes in the sample and the most common haplotype occurring at higher frequency. Additionally, since the SFS treats all SNPs as being independent, haplotype patterns capture more information regarding the local genealogy within a window than the SFS does. In other words, many of the high-frequency derived SNPs in the sample are clustered in certain windows, but this pattern is missed in the SFS since it treats all SNPs as being exchangeable. Notably, summaries of the SFS performed on a local scale may be better at distinguishing these two models.

It is important to point out that the *HCN* statistic has been designed to be used on ascertained SNP data, where the number of SNPs in a particular window of the genome is affected by the ascertainment process. Consequently, we deliberately did not use the number of SNPs in constructing the *HCN* statistic. To analyze full-resequencing data in a haplotype framework, a more powerful approach would also make use of information about the number of SNPs in each window (Innan *et al.* 2005). The *HCN* statistic can be modified to include this information, suggesting that haplotype patterns based on full-resequencing data will be even more informative than described here. Thus haplotype statistics, including the *HCN* statistic, will remain relevant for demographic inference even when ascertainment bias is no longer an issue.

The example above suggests that combining the SFS and *HCN* statistic may present a powerful approach to distinguish between complex demographic scenarios. Further work combining the two statistics for demographic inference is ongoing. Another possible extension of our method would be to jointly model two populations

in an isolation-migration framework (proposed by Nielsen and Wakeley 2001) where the data are summarized by the *HCN* statistic for shared and population specific haplotypes. Finally, instead of using standard coalescent simulations to find the expected *HCN* statistic for a given demographic scenario, we could instead approximate the coalescent using the sequentially Markov coalescent (McVean and Cardin 2005; Marjoram and Wall 2006). Doing so would reduce the computational burden of the method and would also allow for greater values of c_{window} to be used.

CHAPTER 2

COMPARING PATTERNS OF LINKAGE DISEQUILIBRIUM ON THE HUMAN X CHROMSOME AND AUTOSOMES[2]

2.1 Abstract

Population genetic theory predicts that under simple demographic models, the X chromosome in humans should have 3/4 the effective population size as the autosomes. Additionally, the X chromosome differs from the autosomes since it does not recombine in males. For these reasons, levels of linkage disequilibrium (LD) on the X chromosome are predicted to be higher than on the autosomes. Here we use population genetic methods to quantify the amount of LD on the X chromosome and the autosomes. In particular, we use LD patterns to estimate the ratio of the effective population size on the X chromosome (N_X) and on the autosomes (N_{Auto}). Overall, we find that N_X/N_{Auto} is greater than 0.75, indicating that there is too little LD on the X chromosome relative to what is expected based on LD patterns from the autosomes. Using unphased diploid genotypes, we also reject a model where $N_X/N_{Auto} = 0.75$ in regions of low recombination. We discuss several possible explanations for lower than expected levels of LD on the X chromosome, including sex-biased demography, natural selection, and higher levels of gene conversion on the X chromosome than on the autosomes.

[2] Lohmueller, K.E., J.D. Degenhardt, C.D. Bustamante, A.G. Clark, In preparation.

2.2 Introduction

Since the human X chromosome is present in two copies in females, but only in one copy in males, it has population genetic properties that differ from the autosomes. For example, if there are an equal number of breeding males and females in the population, the effective population size on the X chromosome (N_X) is expected to be ¾ that of the autosomes (N_{Auto}), since there are three X chromosomes for every four autosomes contained in every male-female pair (Hedrick 2007). Additionally, it is predicted that natural selection will act differently on the X chromosome than on the autosomes. Beneficial recessive mutations on the X chromosome will be immediately exposed to selection in the hemizygous males, thus allowing selection to operate more efficiently on the X chromosome relative to the autosomes (Charlesworth *et al.* 1987; Begun and Whitley 2000; Singh *et al.* 2008). As a result, more neutral genetic variation linked to a selected site will be removed on the X chromosome than on the autosomes. Similarly, recessive deleterious mutations should be eliminated more efficiently on the X chromosome than the autosomes. Consequently, since fewer neutral mutations will be linked to deleterious mutations on the X chromosome, neutral variation will be reduced less on the X chromosome than on the autosomes due to background selection (Charlesworth *et al.* 1987; Charlesworth *et al.* 1993; Begun and Whitley 2000; Singh *et al.* 2008). Finally, the X chromosome undergoes less recombination on average than the autosomes do (Hedrick 2007). This occurs because the X chromosome only recombines in females, while the autosomes recombine in both males and females.

The differences between the X chromosome and the autosomes discussed above have important implications for comparing patterns of genetic variation on the X chromosome to the autosomes. Namely, the simplest model, assuming strict neutrality and an equal number of breeding males and females predicts, that $N_X/N_{Auto} =$

0.75. Two recent studies have tested whether patterns of genetic variation are consistent with this expectation. Hammer *et al.* (2008) and Keinan *et al.* (2009) compared levels of nucleotide diversity on the X to that on the autosomes. Keinan *et al.* (2009) also compared patterns of population differentiation on the X and autosomes and the frequency spectrum of single nucleotide polymorphisms (SNPs) on the X and the autosomes. Both studies rejected this simple model in non-African populations. However, they disagreed as to whether N_X/N_{Auto} was >0.75 or <0.75. Furthermore, for an African population, Keinan *et al.* (2009) failed to reject a simple model where $N_X/N_{Auto} = 0.75$, while for other African populations, Hammer *et al.* (2008) did reject $N_X/N_{Auto} = 0.75$. It is unclear why the studies disagreed, especially since the analysis of nucleotide diversity in Keinan *et al.* (2009) is similar to the type of analyses used in Hammer *et al.* (2008). However, the two studies considered different genomic regions on the X chromosome. Hammer *et al.* (2008) studied regions of high recombination far away from genes, while Keinan *et al.* (2009) analyzed a random assortment of genomic regions across the chromosome. The differences in the genomic regions surveyed may be partially responsible for the differences in the results.

All the analyses in Hammer *et al.* (2008) and Keinan *et al.* (2009) consider every SNP to be independent of each other, and do not use any information from the correlation patterns of alleles at different SNPs (termed linkage disequilibrium, LD). Patterns of LD have been important in population genetics since they contain information regarding demographic history (Reich *et al.* 2001; Garrigan *et al.* 2005; Conrad *et al.* 2006; Hellenthal *et al.* 2008; Auton *et al.* 2009; Lohmueller *et al.* 2009) as well as how recombination operates across the genome (Reich *et al.* 2002; Wall and Pritchard 2003; Crawford *et al.* 2004; McVean *et al.* 2004a; Ptak *et al.* 2004; International HapMap Consortium 2005; Jeffreys *et al.* 2005; Myers *et al.* 2005; Ptak

et al. 2005; Winckler *et al.* 2005a; Clark *et al.* 2007; International HapMap Consortium 2007). Thus, comparing patterns of LD on the X chromosome to the autosomes may reveal important insights about sex-biased demographic history or different recombinational process on the X chromosome and the autosomes. Since the simple model of an equal number of breeding males and females predicts that $N_X/N_{Auto} = 0.75$, the per-generation population scaled recombination rate, ρ, will be smaller on the X chromosome than on the autosomes. Furthermore, since the X chromosome does not recombine in males, the sex-averaged recombination rate will be smaller on the X chromosome than on the autosomes. Thus, considering both these factors, the simplest model suggests that there should be increased LD on the X chromosome relative to what is seen on the autosomes.

It is unclear whether previous studies of genome-wide LD patterns were consistent with this prediction. For example, Li *et al.* (2008) found smaller estimates of ρ (suggestive of increased LD) on the X chromosome than on the autosomes after normalizing for recombination rate differences. However, these estimates were often larger than ¾ those of the autosomes. Furthermore, Tenesa *et al.* (2007) estimated N_X and N_{Auto} from pairwise patterns of LD. Their estimates of N_X were larger than ¾ those of the autosomes. It is difficult to interpret whether these results are consistent with the simple prediction that $N_X/N_{Auto} = 0.75$, since neither of these studies used population genetic models to formally test such a hypothesis. Additionally, these studies relied on data from genotyping assays without controlling for, or evaluating the effect of, the lower SNP density observed on the X chromosome relative to the autosomes (International HapMap Consortium 2005; International HapMap Consortium 2007).

Here we perform a systematic analysis of patterns of LD on the X chromosome and the autosomes. We summarize LD across genomic regions by the population

scaled recombination rate, ρ, since it provides a quantitative summary of LD that can be interpreted in a population genetic framework. By normalizing for differences in recombination rate, we can then estimate N_X/N_{Auto} for different human populations. We use extensive coalescent simulations accounting for demography, recombination hotspots, SNP ascertainment, and bias in the estimation process to test whether the patterns of ρ are consistent with $N_X/N_{Auto} = 0.75$. For all analyses in both populations studied, we find that N_X/N_{Auto} is greater than 0.75. In some cases, we can reject a model where $N_X/N_{Auto} = 0.75$. We discuss several possible evolutionary explanations for lower than expected levels of LD on the X chromosome.

2.3 Materials and Methods

Analysis of HapMap data

Our first set of analyses used the "consensus" phased haplotypes for the CEU (consisting of individuals from Utah with Northwestern European ancestry) and YRI (consisting of Yoruba individuals from Ibadan, Nigeria) populations from the Phase II HapMap (release 21; International HapMap Consortium 2007). The "consensus" set only includes those SNPs that were assayed in all three HapMap populations. We focused on the CEU and YRI populations since these populations consisted of 30 sets of parents and offspring (trios). Data from the children should allow haplotype phase to be estimated very accurately (Marchini *et al.* 2006). Since our analysis involves comparing LD patterns on the X chromosome to those on the autosomes, and since males have only one X chromosome, we only used the haplotypes from the 30 unrelated females in each of the two populations. Using only females for both the X chromosome and autosomes provides us with the same sample sizes for both analyses, which should provide comparable accuracy in the estimation of ρ for both types of data.

We also repeated our analyses using unphased diploid genotypes taken from the Phase II HapMap (release 23; International HapMap Consortium 2007). Again, we only used genotypes from the 30 unrelated females in each of the two populations to provide similar sample sizes and similar accuracy in the estimation of ρ on the X and autosomes.

We performed two sets of analyses, the first focusing on regions of the genome with high rates of recombination and the second focusing on regions of the genome with lower rates of recombination. Specifically, for the regions of high recombination, we divided the X chromosome and the autosomes into non-overlapping windows of 5 Mb where the average recombination rate for each of these 5 Mb regions was between 5.6-6.4 cM/Mb. For the regions of low recombination, the average recombination rate for each of the 5 Mb regions was between 3.6-4.4. We estimated the average recombination rate for each 5 Mb region using the deCODE genetic map downloaded from the UCSC Genome Browser (Kong *et al.* 2002). Note, we used the sex-averaged recombination rates for the autosomes and the female recombination rate for the X chromosome, since males do not recombine on the X chromosome. We chose to use 5 Mb windows since the deCODE genetic map should have sufficient resolution to provide an accurate estimate of the recombination rate at this scale (Kong *et al.* 2002). We analyzed regions of high and low recombination separately since natural selection has a larger impact on patterns of linked neutral variation in regions of the genome with low recombination rates (Kaplan *et al.* 1989; Begun and Aquadro 1992; Andolfatto and Przeworski 2001; Andolfatto 2001). By analyzing regions of high and low recombination separately, it should be easier, in principle, to assess whether our results are driven by hitchhiking or background selection in regions of reduced recombination.

For each 5 Mb region, we considered SNPs that were polymorphic in the sample of 30 females from both the CEU and YRI populations. The smallest number of SNPs fulfilling this criterion across both X and autosomal windows, denoted *M*, for the regions of high recombination using the phased haplotypes was 1115. Because the number of SNPs within a window will have an impact on the precision of our estimates of ρ (Auton and McVean 2007), we selected M = 1,100 SNPs from each 5 Mb region. Let $P(K_{i,j})$ be the probability of retaining SNP *i* in window *j*. Then

$$P(K_{i,j}) = \frac{M}{L_j}, \qquad (2.1)$$

where L_j is the total number of SNPs in window *j*. Unless otherwise noted, the same set of SNPs were used in both the CEU and YRI populations. However, when comparing our empirical results to simulations that only use single-population demographic models, we independently selected SNPs from the CEU and YRI populations, with the only restriction being that the chosen SNPs were polymorphic within the population of interest, without regard to whether or not they were polymorphic in the other population. We used a similar procedure to select SNPs for the high recombination dataset when performing the analysis on the unphased genotypes, except here we selected 1200 SNPs per region. In the low recombination dataset, we selected 1000 SNPs per region for the phased haplotype analysis and 700 SNPs per region for the analysis using unphased genotypes.

Estimating ρ

We estimated the population scaled recombination rate, ρ, for each 5 Mb region of the genome using the "interval" program in the LDhat package (McVean *et al.* 2004). We used this program since it allows for recombination rate heterogeneity within a region of the genome which is known to occur in humans (McVean *et al.* 2004). We ran LDhat on the CEU and YRI populations separately. We used 1.5 x 10^6

MCMC iterations, with a burn-in of 1 x 10^5 iterations, and retaining one out of every 2000 iterations thereafter. A block penalty of 10 was used. To obtain a point estimate of ρ, we took the median over the 750 values of ρ from the posterior distribution.

Estimating $\hat{N}_X / \hat{N}_{Auto}$ *from* ρ

The population scaled recombination rate, ρ, is determined by both the per-sequence sex-averaged recombination rate and the overall size of the population. This relationship is given by, $\rho = 4N_e r$, where N_e is the effective population size and r is the sex-averaged per-sequence recombination rate. Thus, given the estimate of ρ obtained from the LDhat program as described above, combined with an estimate of the recombination rate from a genetic map, we can obtain an estimate of the population size. Let $\hat{\rho}_{X_i}$ be the estimate of ρ obtained from the LDhat program for the *i*th window on chromosome X and $\hat{\rho}_{Auto_i}$ be the estimate of ρ obtained from the LDhat program for the *i*th window on the autosomes. Then we can calculate $\hat{\rho}_X$ by summing the estimates over all *j* windows on the X chromosome,

$$\hat{\rho}_X = \sum_{i=1}^{j} \hat{\rho}_{X_i} . \qquad (2.2)$$

Let r_{X_i} be the estimate of the total genetic distance for the *i*th window on chromosome X obtained from the sex-averaged deCODE genetic map (Kong *et al.* 2002). Then the total genetic distance on the X chromosome can be calculated by summing over all *j* windows,

$$r_X = \sum_{i=1}^{j} r_{X_i} \qquad (2.3).$$

The same process is used to find $\hat{\rho}_{Auto}$ and r_{Auto}. An estimate of $\hat{N}_X$ is found from

$$\hat{N}_X = \frac{\hat{\rho}_X}{4r_X}, \qquad (2.4)$$

and an estimate of $\hat{N}_{Auto}$ is found from

$$\hat{N}_{Auto} = \frac{\hat{\rho}_{Auto}}{4r_{Auto}} \qquad (2.5).$$

The above conversions require the use of the sex-averaged recombination rates on the X chromosomes and the autosomes. The sex-averaged recombination rate on the autosomes in humans is simply the arithmetic average of the male-and female recombination rates

$$r_{Auto} = \frac{1}{2}\left(r_{Auto\_female} + r_{Auto\_male}\right) \qquad (2.6).$$

Since males do not recombine on the X chromosome, and the X chromosome spends 2/3 of its time in females (regardless of the sex ratio in the population), the sex-averaged recombination rate on the X chromosome is 2/3 the female recombination rate,

$$r_X = \frac{2}{3}\left(r_{X\_female}\right) \qquad (2.7).$$

Further explanation of these formulas can be found in Hedrick (2007).

Coalescent simulations

We undertook a series of coalescent simulations using different demographic models and values of N_X / N_{Auto} to help interpret our estimates of $\hat{N}_X / \hat{N}_{Auto}$ obtained in the YRI and CEU populations. Specifically, we used the coalescent simulations to 1) obtain a bias-corrected estimate of $\hat{N}_X / \hat{N}_{Auto}$, 2) test whether the $\hat{N}_X / \hat{N}_{Auto}$ ratio obtained from the HapMap data were compatible with $N_X / N_{Auto} = 0.75$, and 3) test whether a single value of N_X / N_{Auto} for both the CEU and YRI populations is likely to match the estimates of $\hat{N}_X / \hat{N}_{Auto}$ obtained from these two populations.

Demographic models

We analyzed a total of five demographic models for the CEU and YRI populations. First, the "Schaffner" model considers the CEU and YRI populations

jointly (Schaffner *et al.* 2005). This model also includes migration between Africa and Europe as well as between Africa and East Asia (though we do not include any sampled chromosomes from the East Asian population). This model has been widely used in other population genetic studies and has been found to fit some attributes of empirical data reasonably well (Schaffner *et al.* 2005; Coop *et al.* 2009). By jointly modeling the CEU and YRI population together, this model allows us to test whether the same value of N_X / N_{Auto} in both populations can match the observed values in both populations. The second model used for African demography, termed "SNM", is simply a constant size population of 10,000. The third model used for African demography, termed "Growth", is the growth model that Keinan *et al.* (2007) had fit to the YRI data. This model includes an ancient ~1.8 fold expansion. For the CEU population, we also considered two additional bottleneck models. The first one, termed "Keinan" is the two bottleneck model that Keinan *et al.* (2007) had fit to the CEU data. The final bottleneck model termed "Lohmueller" was a bottleneck model fit to the CEU data using a haplotype-based inference approach (Lohmueller *et al.* 2009).

Simulation strategy

To perform coalescent simulations using the demographic models described above, we used the program *macs* since this program allows rapid simulation of large genomic regions (Chen *et al.* 2009). For each demographic model and value of N_X or N_{Auto}, we simulated 100 datasets. For the X chromosome, each dataset consisted of the same number of 5 Mb regions as in the HapMap data (9 for the high recombination dataset and 7 for the low recombination dataset) with average recombination rates matching the sex-averaged rates for the 5 Mb regions in the HapMap data estimated from the deCODE genetic map. For the autosomes, each dataset consisted of 134 or 110 5 Mb regions, for the high and low recombination datasets, respectively, with

average recombination rates matching the average rates for the regions in the HapMap data estimated from the deCODE genetic map. We then estimated $\hat{N}_X / \hat{N}_{Auto}$ from each simulated dataset by running LDhat on it using the same settings as used on the observed data.

Recombination rate models

The average recombination rates for each simulated 5 Mb region matched the sex-averaged rates from the HapMap data as estimated from the deCODE genetic map. Since recombination rates in humans are known to vary at a fine scale (McVean *et al.* 2004; International HapMap Consortium 2007), we also modeled recombination hotspots. To do this, we used the LDhat-based genetic map as estimated by the International HapMap consortium (International HapMap Consortium *et al.* 2007). This map provided a guide as to the relationship between genetic distance and physical distance at many different points along each 5Mb region. The macs program allows the user to supply a mapping of how the recombination rate changes along the sequence. This feature was used in our simulations.

Matching SNP density and SNP frequencies

Since the SNPs included in the HapMap project are more likely to be at higher frequency than a random set of SNPs sampled from the genome (International HapMap Consortium 2007), we tried to match the frequency spectrum of the SNPs in our coalescent simulations to the frequency spectrum of SNPs in the HapMap data. This was done in one of two ways, depending on whether we were simulating the CEU and YRI populations separately or together. When simulating the two populations separately, for each 5 Mb region in the HapMap data, we tabulated the proportion of SNPs where the minor allele was at frequency 1/60, 2/60, …, 30/60. These proportions were then used in the macs simulation (by using the –F flag) to thin the simulated SNPs such that the frequency distribution for the simulated region

matched the observed frequency distribution for that window. When simulating the two populations together, we used a similar approach as described above, but we considered the frequency distribution over both populations together. In other words, we tabulated the proportion of SNPs at frequency 2/120, 3/120, …, 60/120. The 120 comes from the total number of chromosomes used in both the CEU and YRI samples (60 CEU + 60 YRI). Note, there were no SNPs with minor allele frequency of 1/120 since we only considered SNPs that were polymorphic in both the CEU and YRI populations in the analysis of both the real data and the simulations.

As described above, we thinned the number of SNPs per window in the HapMap data to be the same on the X and the autosomes. We followed a similar process for the simulated data. This was done by using a value the population scaled mutation rate (θ) in the simulations that would give many more than M SNPs, even after filtering SNPs to match the frequency spectrum as described above. Then, similar to what was done in the observed data, we retained each SNP in within a simulated window with probability M/L, where L is the number of SNPs in the simulated window, and M is the desired number of SNPs to be retained in the particular dataset. Note, when modeling the CEU and YRI populations together, we only kept those SNPs that were polymorphic in both populations.

The above strategy should approximately allow the number of SNPs per window and frequency spectrum in the simulations to match those in the observed data. Some subtle differences will exist between the HapMap frequency spectrum and the simulated frequency spectrum since in the simulations, we matched the frequency spectrum using many more than M SNPs, and then dropped SNPs to have M SNPs per window. Dropping SNPs after matching the frequency spectrum may cause the two spectra to match less well. Nevertheless this simulation strategy allows us to better match the HapMap frequency spectrum than would standard coalescent simulations.

Obtaining bias-corrected estimates

Based on application of LDhat to estimate $\hat{N}_X / \hat{N}_{Auto}$ from datasets simulated under a variety of demographic models, we find that the estimates are sometimes biased (i.e. the average of the estimates is not equal to the true N_X / N_{Auto} ratio used to simulate the data). Since the bias can be reproduced in simulations, we can measure its magnitude for different demographic models and input values of N_X / N_{Auto} and then remove the bias in estimates obtained from the HapMap data. To do this, for each demographic model and several different input values of N_X / N_{Auto}, we simulated 100 genome-wide datasets and estimated $\hat{N}_X / \hat{N}_{Auto}$ for each dataset. We then plotted the mean (over the 100 datasets) $\hat{N}_X / \hat{N}_{Auto}$ versus the actual input value used in the simulations (N_X / N_{Auto}). Linear interpolation was used to obtain a bias corrected point estimate of $\hat{N}_X / \hat{N}_{Auto}$ for the CEU and YRI HapMap data.

We obtained 95% confidence intervals (CIs) on our estimates of $\hat{N}_X / \hat{N}_{Auto}$ using non-parametric bootstrapping. This was done by sampling 5 Mb regions with replacement (matching the observed number of regions as in our actual dataset) on the X chromosome and on the autosomes. We then re-calculated $\hat{N}_X / \hat{N}_{Auto}$ for each bootstrap replicate. We then define the CI as $\hat{N}_X / \hat{N}_{Auto} \pm 1.96\sigma$, where $\hat{N}_X / \hat{N}_{Auto}$ is the bias-corrected estimate obtained from the original dataset and σ is the standard deviation of the estimates across bootstrap replicates.

Testing hypotheses

We also used coalescent simulations to test whether our estimates of $\hat{N}_X / \hat{N}_{Auto}$ from the HapMap data would be unusual if the true N_X / N_{Auto} ratio was 0.75. To do this, we simulated 100 genome-wide datasets matching our observed data as described above, assuming that $N_X / N_{Auto} = 0.75$. We then used LDhat to estimate $\hat{N}_X / \hat{N}_{Auto}$ for each of the simulated datasets. We then compared our observed estimates from the HapMap data to the distributions of the estimates from the

simulations. We assessed the fit of the model to the estimated $\hat{N}_X / \hat{N}_{Auto}$ ratio from the HapMap data by calculating two *P*-values: $P_{empirical}$ is simply the fraction of simulation replicates that are further away from the mean of the simulations replicates than the observed estimate, and P_{normal} is the probability that the observed $\hat{N}_X / \hat{N}_{Auto}$ ratio from the HapMap data falls in the extreme tail of a normal distribution with the mean and variance estimated from the simulation replicates. Note, since we are testing the null hypothesis that $N_X / N_{Auto} = 0.75$, all of the *P*-values presented here are from two-sided tests which will reject the null hypothesis if the observed values are too high or too low. The same approach was used to test whether our estimates of $\hat{N}_X / \hat{N}_{Auto}$ from the CEU HapMap data would be unusual if the true N_X / N_{Auto} ratio was 0.635.

We also tested whether the same N_X / N_{Auto} value would be compatible with our observed estimates from the CEU and YRI populations. This was done the same way as described above, except here we compared the estimates from the two populations to the joint distribution of CEU and YRI estimates simulated from the Schaffner demographic model.

2.4 Results

Comparison of LD patterns on the X vs. autosomes

We used LDhat to estimate the population scaled recombination rate, ρ, from the HapMap data. As such, ρ is a summary of LD, where lower estimates of ρ are compatible with higher amounts of LD. Figure 2.1 shows the estimates of ρ obtained in the CEU population vs. the YRI population for 5 Mb regions on the X chromosome (red) and the autosomes (black). For all four datasets shown in Figure 2.1, we note that the estimates of ρ tend to fall below the diagonal on both the X chromosome and the autosomes, suggesting that there is more LD in the CEU population than the YRI population. This finding is expected and is consistent with previous reports (Reich *et*

al. 2001; Ptak *et al.* 2004; Hinds *et al.* 2005; International HapMap Consortium. 2005; Conrad *et al.* 2006; International HapMap Consortium *et al.* 2007; Jakobsson *et al.* 2008; Li *et al.* 2008; Wall *et al.* 2008).

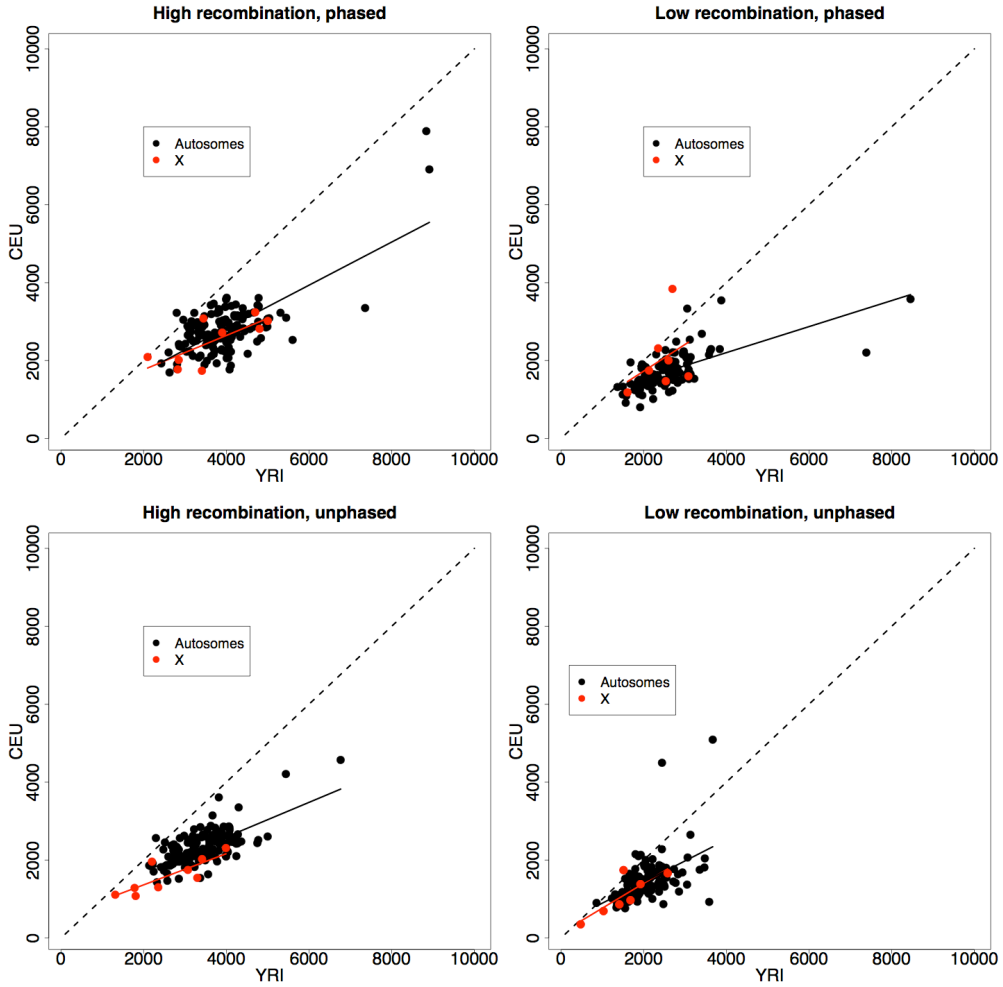


Figure 2.1: Estimates of ρ for the CEU population vs. estimates of ρ in YRI population in the HapMap data. Each panel denotes a different dataset. The dotted line in each panel represents the diagonal, the solid lines represent the best-fit linear regression for the X chromosome (red) and the autosomes (black). Note, overall, estimates of ρ are smaller in CEU than YRI.

Interestingly, when analyzing data from the phased haplotypes (top panels), we find that the estimates of ρ on the X chromosome appear to be within the distribution of the estimates obtained on the autosomes. This observation suggests that LD patterns on the X chromosome are not substantially higher than on the autosomes. When

analyzing the high recombination dataset using unphased genotypes (Figure 2.1, bottom panels), we find that the estimates of ρ tend to be lower on the X chromosome relative to the autosomes.

Estimates of N_X/N_{Auto} from LD patterns

We used the estimates of ρ on the X chromosome and the autosomes to find $\hat{N}_X / \hat{N}_{Auto}$ from the HapMap data (Table 2.1).

Table 2.1: Estimates of $\hat{N}_X / \hat{N}_{Auto}$ from the HapMap data.

Pop.	Rec. Rate[1]	Phase[2]	Model[3]	Uncorrected $\hat{N}_X / \hat{N}_{Auto}$	Bias-corrected $\hat{N}_X / \hat{N}_{Auto}$	CI low	CI high
YRI	High	Phased	Schaffner	1.395	1.391	1.145	1.637
			SNN	1.323	1.283	0.998	1.569
			Growth	1.323	1.293	1.008	1.578
		Unphased	Schaffner	1.133	0.851	0.603	1.099
	Low	Phased	Schaffner	1.408	1.648	1.434	1.862
		Unphased	Schaffner	1.082	0.924	0.635	1.208
CEU	High	Phased	Schaffner	1.340	1.490	1.285	1.695
			Lohmueller	1.275	1.567	1.358	1.775
			Keinan	1.275	1.362	1.153	1.571
	High	Unphased	Schaffner	1.026	0.789	0.614	0.964
	Low	Phased	Schaffner	1.744	2.649	2.131	3.149
		Unphased	Schaffner	1.099	1.117	0.806	1.428

[1]"High" refers to the dataset consisting of regions with a high recombination rate and "Low" refers to the dataset consisting of regions with a low recombination rate.
[2]"Phased" indicates that the inference was run on phased haplotypes while "Unphased" indicates that the inference was done on unphased genotypes.
[3]The demographic model under which the simulations were done to obtain the bias-corrected estimates (see Methods).

As discussed previously, since the estimates of $\hat{N}_X / \hat{N}_{Auto}$ were sometimes biased, we use coalescent simulations under a variety of demographic models to quantify the bias which is then subtracted from the estimates. Figure 2.2 shows an example of this bias for data simulated under the Schaffner demographic model.

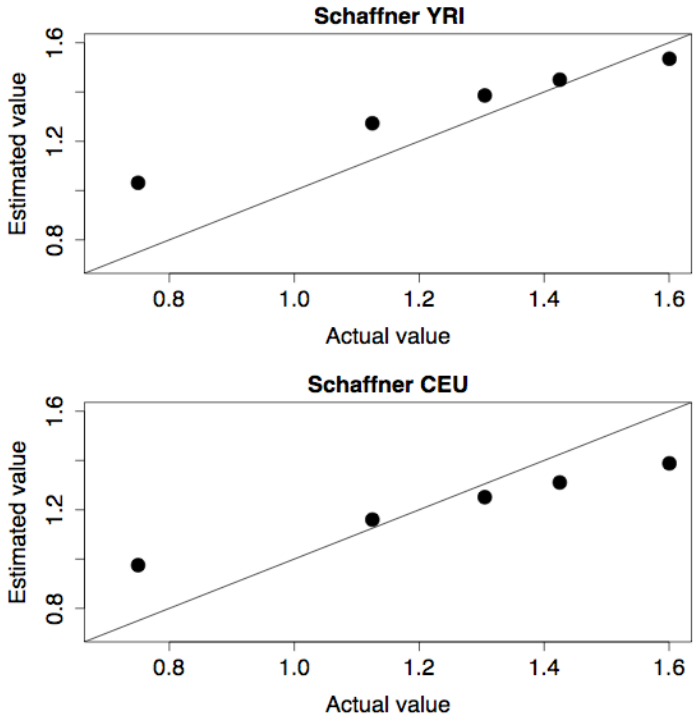


Figure 2.2: Bias in the estimates of $\hat{N}_X / \hat{N}_{Auto}$. Here we plot the average estimate of the $\hat{N}_X / \hat{N}_{Auto}$ ratio in simulated datasets vs. the true value used to simulate the data under the Schaffner demographic model using phased haplotypes and matching the recombination rates to those from the high recombination dataset. The solid line in each panel represents the diagonal. Unbiased estimates of $\hat{N}_X / \hat{N}_{Auto}$ would fall on this line.

Ideally, the average, across simulation replicates, of the estimates of $\hat{N}_X / \hat{N}_{Auto}$ should be equal to the value of N_X / N_{Auto} used to simulate the data. In other words, the points in Figure 2.2 should all lie along the diagonal. Instead, for lower values of N_X / N_{Auto} (e.g. $N_X / N_{Auto} = 0.75$), the estimates appear to be too large and for higher N_X / N_{Auto} values (e.g. $N_X / N_{Auto} > 1.4$), the estimates become downwardly biased. Thus, it appears that the estimates of $\hat{N}_X / \hat{N}_{Auto}$ cannot take on the full range of values used to simulate the data and suggests that for lower rates of recombination, estimates of ρ are upwardly biased, and for higher rates, estimates of ρ are downwardly biased. It is also worth noting that the bias of the estimates differs between the two demographic models. The YRI demographic model includes population growth and a

very slight bottleneck (Schaffner *et al.* 2005). The CEU demographic model includes a more severe bottleneck (Schaffner *et al.* 2005). Thus, as has been noted for nucleotide diversity (Pool and Nielsen 2007), demographic departures from the standard neutral model can affect estimates of the $\hat{N}_X / \hat{N}_{Auto}$ ratio estimated in our approach. Importantly, the bias-corrected estimates of the $\hat{N}_X / \hat{N}_{Auto}$ ratio shown in Table 2.1 remove both the inherent biases in the estimation process as well as biases that result from the departures from the standard neutral demographic model. An additional consequence of this bias correction is the bias-corrected estimates obtained from the CEU and YRI populations should be similar to each other if the true underlying N_X / N_{Auto} ratio is the same in these two populations.

For the high recombination regions, using phased haplotypes, we find that the bias-corrected value (using the Schaffner demographic model) of $\hat{N}_X / \hat{N}_{Auto}$ is 1.39 (1.14-1.64) for the YRI and 1.49 (1.28-1.70) for the CEU. The confidence intervals here do not contain $N_X / N_{Auto} = 0.75$, suggesting that the observed estimate of $\hat{N}_X / \hat{N}_{Auto}$ is not consistent with 0.75. When using other demographic models to perform the bias-correction (see Methods), we obtain similar results, with the bias corrected point estimates ranging from 1.28 to 1.39 in YRI and 1.36-1.57 in CEU. These estimates of $\hat{N}_X / \hat{N}_{Auto}$ appear to be much higher than the point estimates obtained in previous studies. For example, Hammer et al.'s highest point estimate was 1.08 in the Basque population (Hammer *et al.* 2008), and Keinan *et al.* (2009) found values ≈ 0.75 in the YRI and smaller value in the non-African populations.

We also estimated $\hat{N}_X / \hat{N}_{Auto}$ from the high recombination dataset using unphased genotypes, rather than phased haplotypes. For this analysis, the bias-corrected points estimates of $\hat{N}_X / \hat{N}_{Auto}$ that are higher than 0.75, (0.85 in the YRI and 0.79 in the CEU), however, the 95% CIs do include 0.75 (Table 2.1). Here the bias uncorrected and bias corrected point estimates obtained from the YRI sample are

higher than that from the CEU population, although the difference is slight. Interestingly, the 95% CIs on the estimates from both populations do not overlap with the estimates made using phased haplotypes (see Discussion).

We next estimated $\hat{N}_X / \hat{N}_{Auto}$ from the low recombination dataset using phased haplotypes. Again, the bias-corrected point estimates of $\hat{N}_X / \hat{N}_{Auto}$ are higher than 0.75 (2.65 in CEU and 1.65 in YRI). For both populations the 95% CIs do not include 0.75. Interestingly, here $\hat{N}_X / \hat{N}_{Auto}$ in CEU is higher than in the YRI, and the 95% CIs do not overlap. The CIs for $\hat{N}_X / \hat{N}_{Auto}$ in the low recombination dataset in the YRI population overlap with those from the YRI high recombination dataset when performing inference using the phased haplotypes. For the CEU, however, the 95% CIs from the low recombination rate regions do not overlap with those from the high recombination rate regions. There appears to be one region on the X chromosome with an usually large estimate of ρ (and consequently N_X) in the CEU population which may be responsible for this pattern (Figure 2.1).

Finally, we estimated $\hat{N}_X / \hat{N}_{Auto}$ from the low recombination dataset using unphased genotypes. The bias-corrected point estimates of $\hat{N}_X / \hat{N}_{Auto}$ are again higher than 0.75 in both populations (1.117 in CEU and 0.924 in YRI). In the CEU, the 95% CI excludes 0.75, suggesting that the ratio is higher than expected in the CEU population. For both populations, when conducting the inference on the unphased genotypes, the 95% CIs on the estimates from the low and high recombination datasets overlap with each other, suggesting that the estimates are compatible with one another. Again, however, we note that the estimates on the low recombination rate regions using the unphased genotypes are substantially lower than those when using the phased haplotypes. The 95% CIs from the analyses using phased haplotypes do not overlap with those from the unphased genotypes, suggesting that there is a pronounced difference between whether the inference is done on phased or unphased data.

Can a model where $N_X / N_{Auto} = 0.75$ fit the data?

We next used coalescent simulations to assess whether our observed estimates of $\hat{N}_X / \hat{N}_{Auto}$ were compatible with a true value of $N_X/N_{Auto} = 0.75$, under different demographic models. Figure 2.3 shows the distribution of $\hat{N}_X / \hat{N}_{Auto}$ estimates obtained from the high recombination dataset using phased haplotypes. The observed estimates of $\hat{N}_X / \hat{N}_{Auto}$ from the HapMap data are higher than all of the simulated estimates for both populations and for all demographic models tested ($P_{empirical} < 0.01$; $P_{normal} < 0.01$ for all datasets and models in Figure 2.3). These results strongly suggest that the phased haplotypes from the HapMap data are not compatible with the simple model of $N_X/N_{auto} = 0.75$, even under a variety of demographic models.

Figure 2.4 shows a similar analysis for the high recombination dataset when the inference was done using unphased genotypes. The observed $\hat{N}_X / \hat{N}_{Auto}$ estimate from the YRI HapMap data is higher than the mean of the simulated distribution where $N_X/N_{Auto} = 0.75$, but it does not fall in the extreme tail ($P_{empirical} = 0.07$, $P_{normal} = 0.08$). For the CEU data, the observed estimate from the HapMap data falls well within the estimates from the simulated data ($P_{empirical} = 0.67$, $P_{normal} = 0.63$). Thus, for both populations, the observed estimates of $\hat{N}_X / \hat{N}_{Auto}$ are somewhat consistent with 0.75, in contrast to what was found when performing inference on the phased haplotypes. Possible reasons for this discrepancy are discussed below.

Figure 2.5 shows the same analysis for the low recombination dataset when performing inference on the unphased haplotypes. For both populations the observed $\hat{N}_X / \hat{N}_{Auto}$ estimates from the HapMap data are higher than expected under the Schaffner demographic model if $N_X/N_{Auto} = 0.75$ (YRI: $P_{empirical} = 0.03$, $P_{normal} = 0.033$; CEU: $P_{empirical} < 0.01$, $P_{normal} = 0.012$). Thus, the low recombination rate regions are not compatible with a model where $N_X/N_{Auto} = 0.75$.

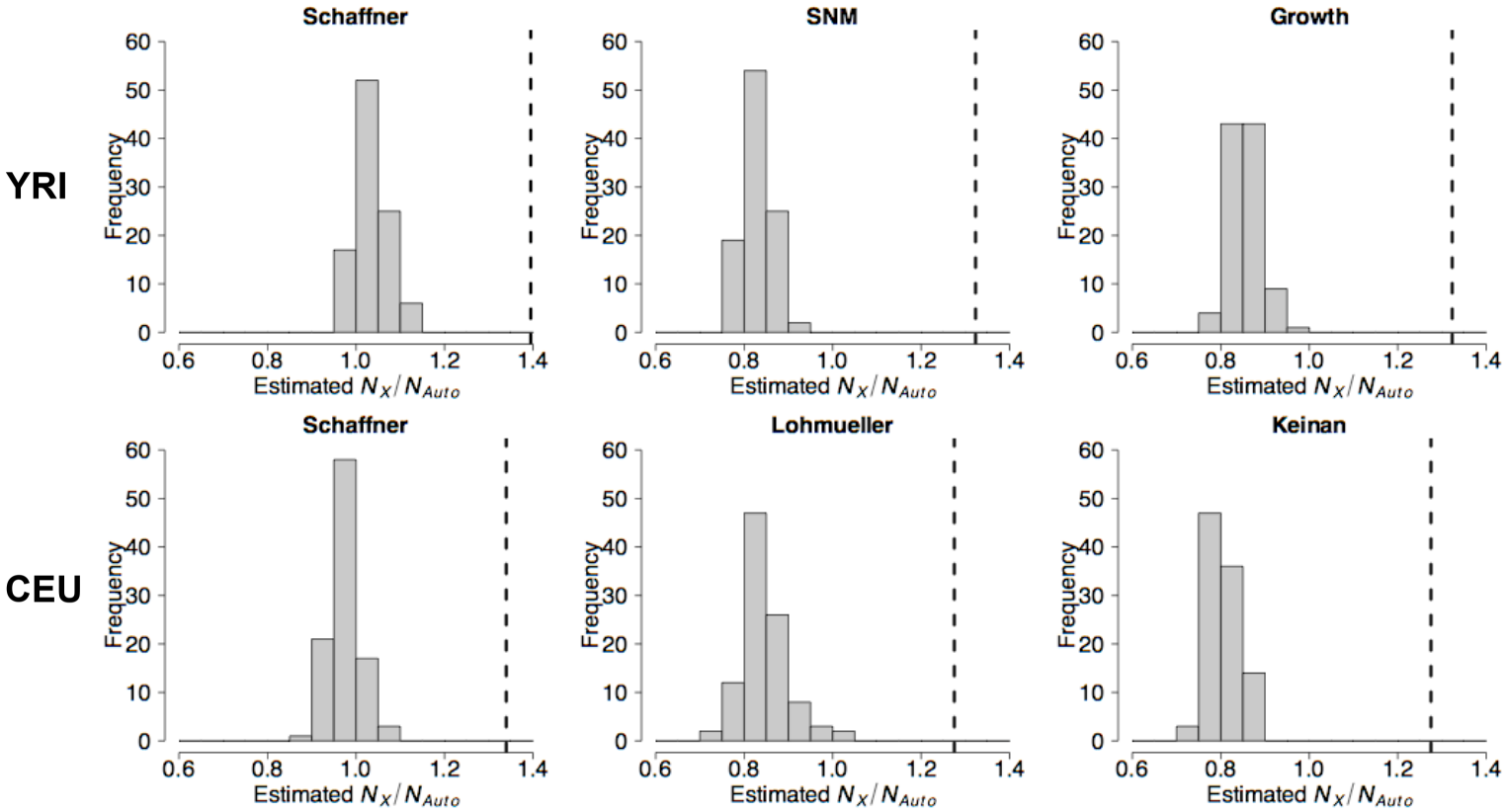


Figure 2.3: Distribution of estimates when the true N_X/N_{Atuo} ratio is 0.75 estimated using phased haplotypes matching recombination rates (cM/Mb) to the high recombination dataset under a variety of demographic models. The vertical dashed line shows the observed estimate from the HapMap data. Note, these are the bias un-corrected estimates. Schaffner denotes the Schaffner demographic model where the CEU and YRI populations were modeled jointly. The other models separately consider CEU and YRI: SNN denotes the standard neutral model, Growth a population expansion model, Lohmueller and Keinen refer to two different bottleneck models that have previously been fit to the CEU data. See Methods for further details.

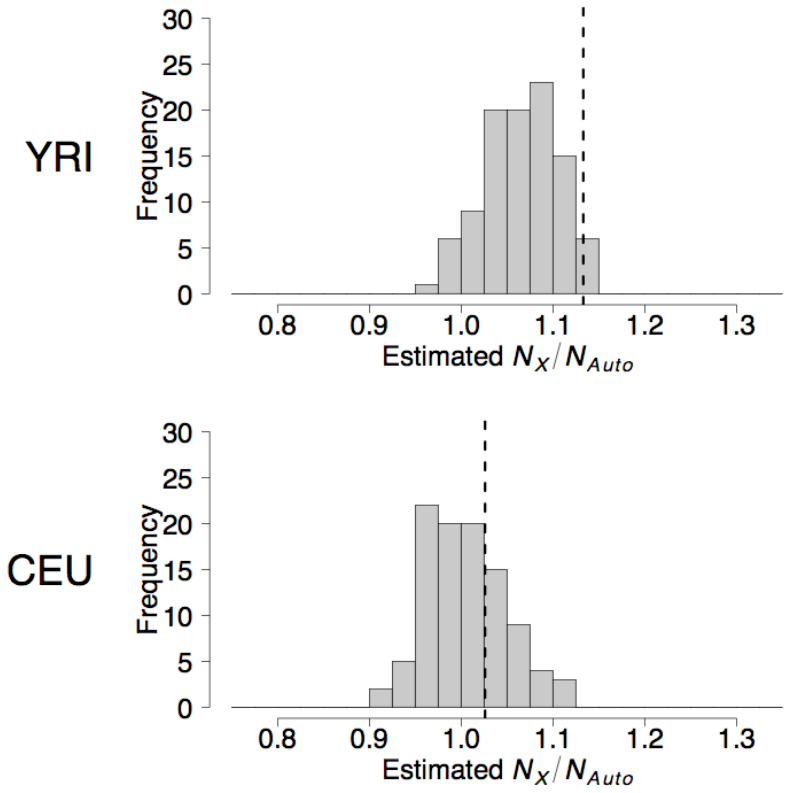


Figure 2.4: Distribution of $\hat{N}_X / \hat{N}_{Auto}$ estimates when the true N_X / N_{Auto} ratio is 0.75 estimated using unphased genotypes matching recombination rates (cM/Mb) to the high recombination dataset under the Schaffner demographic model. The vertical dashed line shows the observed estimate from the HapMap data. Note, these are the bias un-corrected estimates. See Methods for further details.

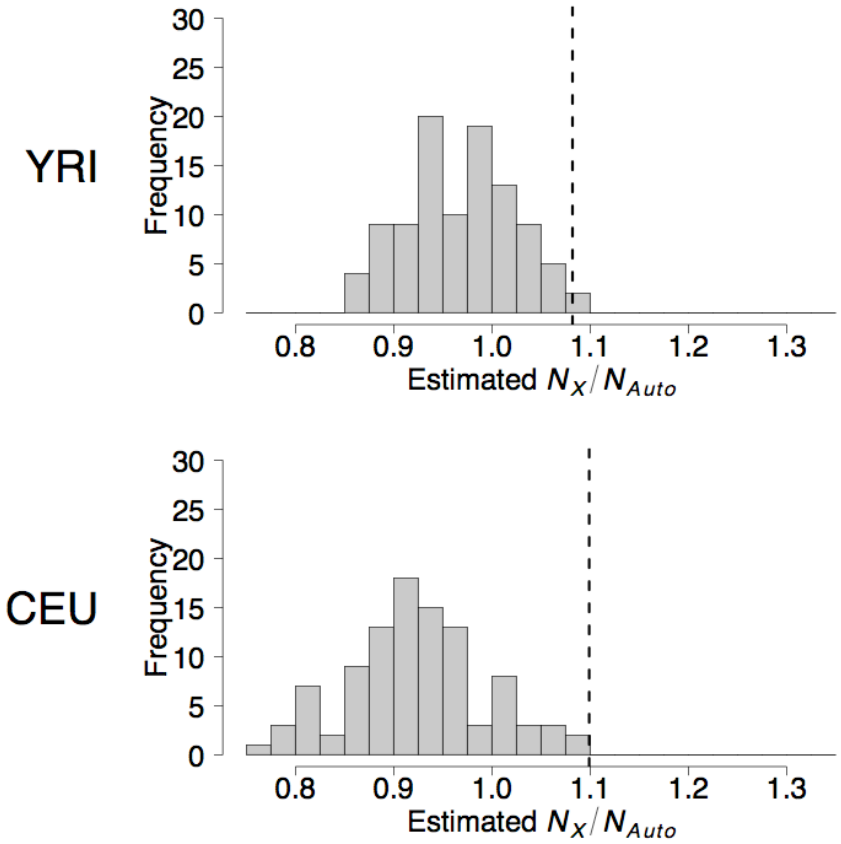


Figure 2.5: Distribution of $\hat{N}_X / \hat{N}_{Auto}$ estimates when the true N_X / N_{Auto} ratio is 0.75 estimated using unphased genotypes matching recombination rates (cM/Mb) to the low recombination dataset under the Schaffner demographic model. The vertical dashed line shows the observed estimate from the HapMap data. Note, these are the bias un-corrected estimates. See Methods for further details.

Can a model where $N_X / N_{Auto} = 0.635$ fit the CEU data?

Keinan *et al.* (2009) used the normalized average number of pairwise differences to estimate that N_X / N_{Auto} was 0.635 in the CEU. Here we test whether the estimates of $\hat{N}_X / \hat{N}_{Auto}$ derived from LD information are compatible with this result. Figure 2.6 shows that using the unphased genotypes, our observed $\hat{N}_X / \hat{N}_{Auto}$ estimates from the CEU population are unlikely under a model the true N_X / N_{Auto} ratio is 0.635 for both the high ($P_{empirical} = 0.02$; $P_{normal} = 0.025$; Figure 2.6A) and low ($P_{empirical} < 0.01$; $P_{normal} < 0.004$; Figure 2.6B) recombination datasets.

The same N_X / N_{Auto} value fits both CEU and YRI

So far we have analyzed the YRI and CEU populations separately. We next asked whether the same value of N_X/N_{Auto} could be compatible with the estimates of $\hat{N}_X / \hat{N}_{Auto}$ we found in both populations in the HapMap data. This analysis was motivated by the study of Keinan *et al*. who found that the same value of N_X/N_{Auto} could not fit both the CEU and YRI populations. Figure 2.7A shows a scatterplot of the estimates of $\hat{N}_X / \hat{N}_{Auto}$ obtained from the 100 simulated datasets under the Schaffner demographic model assuming that $N_X/N_{Auto} = 0.75$. The circle denotes the 95% confidence region assuming that the estimates from the simulated datasets follow a bivariate normal distribution. The observed estimate of $\hat{N}_X / \hat{N}_{Auto}$ from the HapMap high recombination data estimated from unphased genotypes (the red dot) falls well within this region, suggesting that we cannot reject a model where the N_X/N_{Auto} ratio is the same in both populations. Figure 2.7B shows the same results for the low recombination rate regions, although here N_X/N_{Auto} used to simulate the data is 1.125, instead of 0.75.

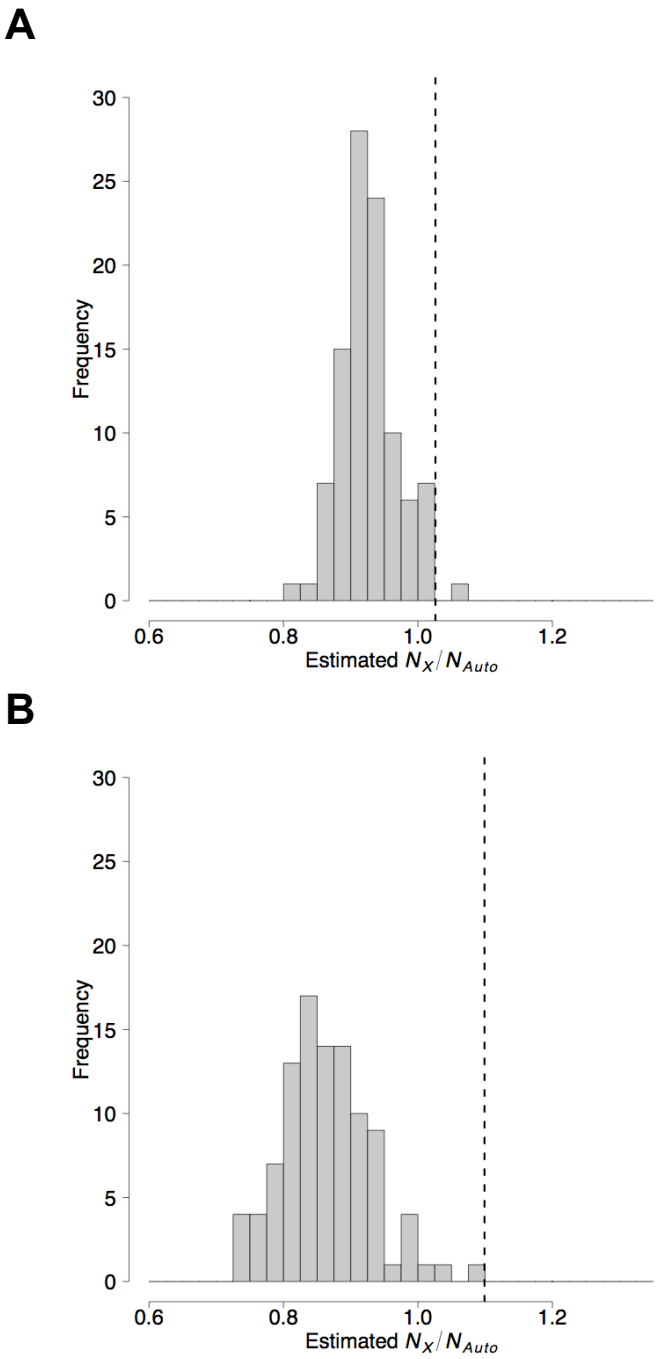


Figure 2.6: Distribution of $\hat{N}_X / \hat{N}_{Auto}$ estimates when the true N_X / N_{Auto} ratio is 0.635 estimated using unphased genotypes matching recombination rates (cM/Mb) to the high (A) and low (B) recombination dataset under the Schaffner demographic model. The vertical dashed line shows the observed estimate from the HapMap data. Note, these are the bias un-corrected estimates. See Methods for further details.

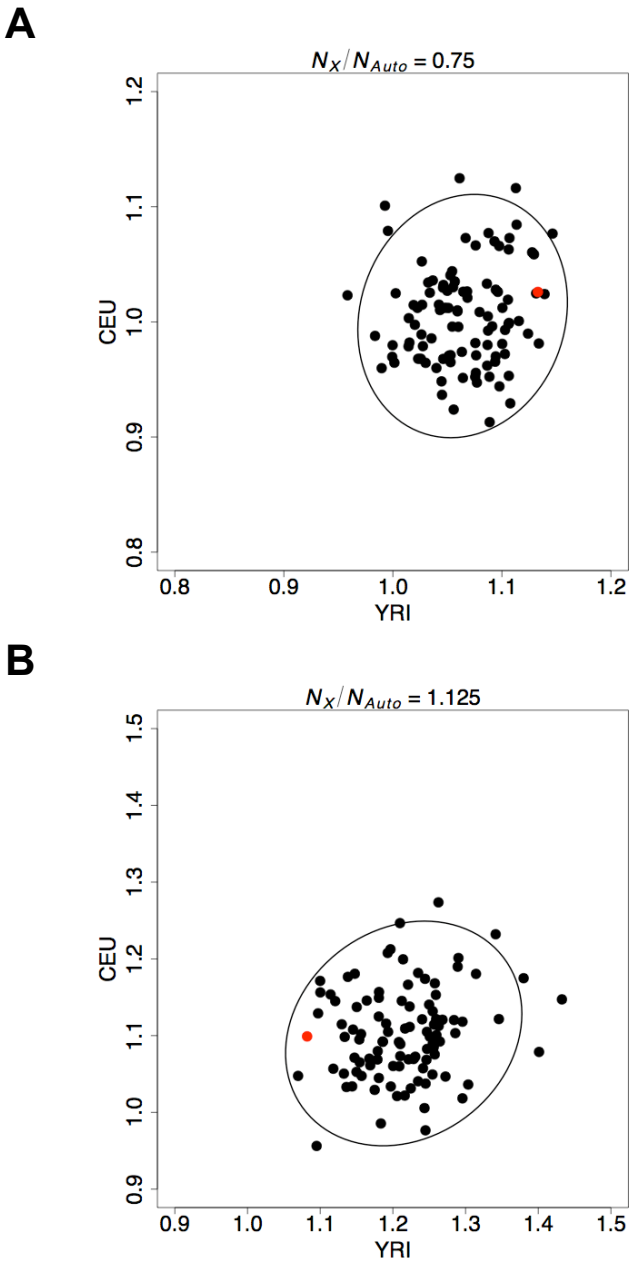


Figure 2.7: Joint distribution of the $\hat{N}_X / \hat{N}_{Auto}$ estimates from the simulated CEU and YRI populations under the Schaffner demographic model (black points). Note, here unphased genotypes were used and the recombination rates in the simulations were matched to the high recombination rate dataset (A) and the low recombination dataset (B). The true N_X / N_{Auto} ratio used to simulate the data was 0.75 in A and 1.123 in B. The red dot denotes the observed estimate from the appropriate dataset using unphased genotypes from the HapMap data. The ellipse shows the 95% confidence region obtained from a bivariate normal distribution fit to the simulated estimates.

2.5 Discussion

There are several noteworthy aspects regarding estimates of $\hat{N}_X / \hat{N}_{Auto}$ for the CEU and YRI populations based on patterns of LD. First, the estimates of $\hat{N}_X / \hat{N}_{Auto}$ when using phased haplotypes are significantly higher than when using unphased genotype data. Below we explain why this finding is likely due to a higher rate of phasing errors on the X chromosome than on the autosomes. As such, our inferences performed using the unphased genotype data are likely to be more robust and these inferences are the ones discussed in the remainder of this section. Even when analyzing the unphased genotypes, we still find evidence that $\hat{N}_X / \hat{N}_{Auto} > 0.75$. We next discuss and evaluate several explanations for why $\hat{N}_X / \hat{N}_{Auto}$ could be >0.75 evaluate which of these explanations can reconcile our estimates with those obtained by Keinan *et al.* (2009) and Hammer *et al.* (2008).

Reconciliation of estimates of $\hat{N}_X / \hat{N}_{Auto}$ from phased haplotypes vs. unphased genotypes

Our estimates of $\hat{N}_X / \hat{N}_{Auto}$ made using phased haplotypes are significantly larger than those made using unphased genotypes. One potential explanation for this result is that LDhat provides more accurate estimates on phased haplotypes, rather than diploid genotypes. However, our simulation studies (Figures 2.4 and 2.5) suggest that LDhat performs well even on diploid genotypes. Furthermore, if differences in the performance of LDhat on phased and unphased data can explain this difference, the bias-correction based on the simulations should remove this effect. Thus, it does not seem that a simple difference in the amount of information contained in phased haplotypes versus genotypes, or differences in the performance of LDhat on the different types of data can explain our observed discrepancy.

In principle, the phased haplotypes obtained from the HapMap project should be accurate since the availability of trio data substantially increases the accuracy of the

phasing. Furthermore, since the male parent of each trio can have his phase unambiguously determined, it should be easier to phase the X chromosome than the autosomes. However, this appears to not be the case. Kidd *et al.* (2008) assessed the accuracy of the HapMap phased haplotypes by comparing the haplotypes to fosmid end-sequence pairs aligned to the human reference genome (Kidd *et al.* 2008). For the autosomes, they found that the phased haplotypes provided by the International HapMap Consortium were highly accurate. For example, 0.82% of autosomal sites in individual NA19129 (a child from a YRI trio) covered by clones from fosmid libraries were discrepant with the inferred HapMap haplotypes (Kidd *et al.* 2008). The picture was entirely different for the X chromosome. Here, Kidd *et al.* (2008) found 10 times more phasing errors on the X than on the autosomes. For example, 13.49% of sites in individual NA1929 covered by cones from the fosmid libraries were discrepant with the inferred haplotypes (J. Kidd, personal communication). The analysis by Kidd *et al.* (2008) used the same version of X chromosome haplotypes as we did for the present study (release 21, posted on the HapMap website in May 2007). Apparently these defective files have not been corrected and are still available for download. Curiously, later releases of the phase II HapMap (release 22 and 23) and the HapMap III data do not contain phased haplotypes from chromosome X.

Since the accuracy of the phased haplotypes from chromosome X is quite suspect, our analyses based on the unphased genotype are more reliable. Indeed, our simulations suggest that accurately phased haplotypes should give comparable estimates of $\hat{N}_X / \hat{N}_{Auto}$ to unphased genotypes. Thus, for the remainder of the Discussion, we only interpret the results of the analyses using unphased genotypes.

Explanations for elevated $\hat{N}_X / \hat{N}_{Auto}$

We find that $\hat{N}_X / \hat{N}_{Auto} > 0.75$ for both populations (Table 2.1). For the low recombination regions, we also reject a model where $N_X / N_{Auto} = 0.75$ (Figure 2.6).

Below we discuss and evaluate several possible explanations for this pattern: 1) further systematic biases in HapMap data, 2) sex-biased demographic effects, 3) differences in natural selection on the X vs. autosomes, 4) differences in hotspot usage on the X and autosomes, and 5) differences in gene conversion patterns on the X vs. autosomes.

Systematic biases in the HapMap data

One concern is that our analyses described above use data from the HapMap project which is known to have been influenced by ascertainment bias (Clark *et al.* 2005; International HapMap Consortium 2007). Since SNPs are often discovered in small sample sizes, the HapMap project is enriched for common SNPs over rare SNPs, although the phase II HapMap is less biased than the phase I HapMap (International HapMap Consortium 2007). Additionally, the decision to type some SNPs in HapMap was based on LD patterns between those SNPs and other SNPs (International HapMap Consortium 2007). If these biases were systematically different between the X chromosome and the autosomes, in principle, they could contribute to our high estimates of $\hat{N}_X / \hat{N}_{Auto}$. Short of having full-resequencing data, it is hard to directly evaluate whether ascertainment bias has a meaningful effect on our inferences. However, we note that the HapMap data have been extensively used for estimating both fine-scale and broad scale recombination patterns across the genome by using LDhat, suggesting that our analyses are appropriate for this type of data (International HapMap Consortium 2007).

As an additional quality control measure, we repeated our analysis on the CEU sample from the Perlegen genome-wide polymorphism dataset (Hinds *et al.* 2005). This analysis considered data from the 11 CEU females and was done on the same regions as those in our high recombination dataset. We ran LDhat using unphased genotypes and here we selected 300 SNPs per 5 Mb window. Importantly, we

restricted our analysis to the "Type A" SNPs which were discovered by array-based resequencing of a multi-ethnic panel. The "Type A" SNPs represent a more uniform ascertainment of SNPs across the genome (Clark *et al.* 2005; Hinds *et al.* 2005). The Perlegen data shows $\hat{N}_X / \hat{N}_{Auto} = 1.376$ which is substantially greater than 0.75. While the SNP density and number of individuals studied in the Perlegen dataset is low, finding an elevated $\hat{N}_X / \hat{N}_{Auto}$ ratio when using uniformly ascertained SNPs suggests that are results from the HapMap data are unlikely to be completely driven by ascertainment biases in the HapMap data.

Sex-biased demographic effects

The elevated $\hat{N}_X / \hat{N}_{Auto}$ ratio found in the HapMap data could also be due to demography. A larger female than male effective population size could explain our results. If the variance in reproductive success is greater in males than in females, then females would have a larger effective population size than males (Hammer *et al.* 2008).

Such a demographic explanation was favored by Hammer *et al.* (2008) to explain their estimates of the N_X / N_{Auto} ratio that were higher than 0.75. Their estimates of $\hat{N}_X / \hat{N}_{Auto}$ range from ~0.85 to ~1.05 and appear to be well-within the ranges of our estimates from both populations made from the unphased genotypes (Table 2.1). Thus, if the differences in LD patterns between the X and the autosomes can be explained solely by demographic effects, then our results are consistent with those of Hammer *et al.* (2008). Under such a scenario, our results are then inconsistent with those of Keinan *et al.* (2009) who found that $\hat{N}_X / \hat{N}_{Auto}$ was <0.75 in CEU and equal to 0.75 in YRI and that a single value of N_X / N_{Auto} could not explain both populations. Based on our analysis using patterns of LD, a single value of N_X / N_{Auto} could fit both populations (Figure 2.7), although since our CIs are quite large, we may not have power to detect slight differences. For both the high and low

recombination regions, we were able to reject the point estimate for the N_X / N_{Auto} ratio in CEU (0.635) obtained by Keinan *et al.* (2009; Figure 2.6). However, our model assumed that $N_X / N_{Auto} = 0.635$ throughout all of evolutionary history, which is unlikely. Keinan *et al.* (2009) suggested a model where $N_X / N_{Auto} = 0.75$ throughout much of history, except soon after the Out of Africa bottleneck where multiple waves of male migration occurred. Thus, it is entirely possible that such a complicated demographic scenario could have a different impact on nucleotide diversity and LD patterns and we would be unable to reject such a model. Nevertheless, such a model would be unable to reconcile the fact that we reject a model where $N_X / N_{Auto} = 0.75$ in YRI and Keinan *et al.* (2009) do not.

Differences in natural selection on the X vs. autosomes

Due to the fact that males are hemizygous for the X chromosome, the dynamics of natural selection on recessive mutations will differ on the X chromosome from the autosomes (Charlesworth *et al.* 1987). Genetic hitchhiking is expected to lead to a greater reduction of linked neutral diversity on the X chromosome than on the autosomes. Indeed, selective sweeps have been offered as an explanation for why X chromosomal diversity levels were below those predicted from the autosomal diversity levels in *Drosophila melanogaster* (Aquadro *et al.* 1994) and *Drosophila simulans* (Begun and Whitley 2000). Furthermore, there is suggestive evidence that the human X chromosome contains a higher fraction of genes evolving adaptively (Chimpanzee Sequencing and Analysis Consortium 2005; Nielsen *et al.* 2005). Thus, for both these reasons, in principle, natural selection could lead to differences in patterns of variation on the X chromosome relative to the autosomes.

It is unclear, however, how selective sweeps could explain our estimates of $\hat{N}_X / \hat{N}_{Auto}$ based upon LD patterns. Simulations have shown that selective sweeps should either have little effect on estimates of ρ (McVean 2007) or tend to decrease

estimates of ρ, rather than increase them (O'Reilly *et al.* 2008). Thus, for selective sweeps to explain our observations, there would have to be disproportionately fewer sweeps on chromosome X relative to the autosomes. This direction is opposite of what would be expected and there is no current evidence to suggest that this is the case. However, the simulations described above were for the situation where a single selective sweep resulted in the fixation of a co-dominant mutation (McVean 2007; O'Reilly *et al.* 2008). Recurrent selective sweeps and/or selection on acting on recessive alleles may have a different impact on LD patterns than predicted by the simple models (Przeworski 2002). However, even under more complicated models, selection would still be predicted to increase LD, with the magnitude of the decrease dependent on the specifics of the selection.

The effect of background selection on ρ has not been studied using simulations, so it is harder to predict whether background selection could explain our results. In principle, the same argument used to explain why background selection may be expected to result in an increase in diversity on the X chromosome could be used to for estimates of ρ. Since background selection will eliminate partially recessive strongly deleterious mutations more efficiently on the X chromosome, a greater fraction of gametes in the population remain free of deleterious mutations. Thus, the overall reduction in population size due to background selection on the X chromosome will not be as great as on the autosomes, leading to higher estimates of ρ on the X chromosome than the autosomes (Charlesworth *et al.* 1993; Aquadro *et al.* 1994). In principle, this could be an explanation for our results. However, the above discussion assumes that background selection can be modeled simply as a reduction in population size. Since this approach does not correctly describe the dynamics of weakly deleterious mutations, it may not be directly applicable to many of the negatively selected mutations which have been found to be weakly deleterious (Boyko

et al. 2008). Because the effect of weakly deleterious mutations on estimates of ρ has not been evaluated, it is not possible to conclusively eliminate this effect from contributing to the elevated estimates of $\hat{N}_X / \hat{N}_{Auto}$ measured from patterns of LD.

Differences in hotspot usage on the X and autosomes

Previous analyses of the HapMap data have estimated fine-scale recombination rates across the genome (International HapMap Consortium 2005; International HapMap Consortium 2007). Interestingly, one study (International HapMap Consortium 2007) found that a slightly higher fraction of recombination tends to occur in a smaller proportion of the sequence on the X chromosomes relative to the autosomes. In other words, more of the recombination on the X chromosome is clustered into hotspots than on the autosomes.

Differences in fine-scale recombination patterns of the X chromosome and the autosomes may be worrying because simulations studies have shown that the performance of LDhat at estimating average recombination rates is somewhat dependent on the fine scale recombination structure of the region in question (Auton and McVean 2007). More accurate estimates of average values of ρ have been obtained from simulated data with constant recombination rates as opposed to simulated data containing recombination hotspots. However, we attempted to mitigate this problem by including a model of recombination rate variation in our coalescent simulations used to test whether the data were compatible with certain demographic models. In particular we used the fine-scale genetic map estimated using LDhat from the HapMap data as a guide to how recombination rates varied at fine scales across all regions studied (International HapMap Consortium 2007). If our results could be explained by differences in the performance of LDhat on the X chromosome and autosomes due to the differences in the fine-scale genetic map, then we would expect to see the same elevated $\hat{N}_X / \hat{N}_{Auto}$ estimates from the simulated datasets. We do not

see such a pattern in the simulated data. Even when accounting for the higher fraction of recombination which occurs in hotspots in the X chromosome, when considering the low recombination regions, we still can reject a model where $N_X/N_{Auto} = 0.75$.

Differences in gene conversion patterns on the X vs. autosomes

We used the deCODE genetic map to convert our estimates of ρ_X and ρ_{Auto} into estimates of N_X and N_{Auto} which should normalize for differences in recombination rates across genomic regions. Since genetic maps from pedigrees contain a limited number of genetic markers fairly far apart, the recombination rates estimated from them are reflective of the crossover rates and are not influenced by gene conversion (Andolfatto and Nordborg 1998; Przeworski and Wall 2001). Thus, our normalization using the deCODE map will not account for differences in gene conversion patterns on the X chromosome and autosomes. Increased gene conversion on the X chromosome relative to the autosomes could in principle explain our results since it has been previously shown that gene conversion can decrease LD (Ardlie *et al.* 2001; Frisse *et al.* 2001; Przeworski and Wall 2001).

However, because these studies which had examined the effect of gene conversion on LD patterns often considered different summaries of LD, marker densities, and sizes of genomic regions, we performed additional simulations to investigate the impact of gene conversion on estimates of ρ using LDhat when using the same type of dataset considered here. We used LDhat to estimate N_X / N_{Auto} from simulated datasets assuming the Schaffner demographic model where the inference was performed on the unphased genotypes. The first set of simulated datasets only included crossovers on both the X chromosome and the autosomes (gray bars in Figure 2.8). The second simulated datasets included the same rate of crossovers as the first, but also included gene conversion only on the X chromosome (black bars in Figure 2.8). The simulations assume that gene conversion events occur at 5 times the

rate of crossover events and that the mean gene conversion tract length is 500 bp, consistent with previous estimates from human data (Ardlie *et al.* 2001; Frisse *et al.* 2001; Padhukasahasram *et al.* 2004; Ptak *et al.* 2004; Padhukasahasram *et al.* 2006).

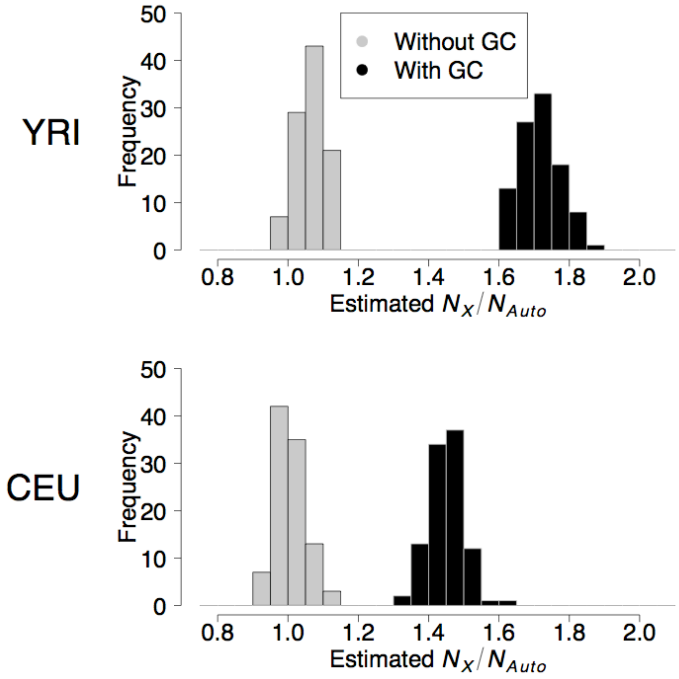


Figure 2.8: Higher rates of gene conversion on the X chromosome relative to the autosomes can give higher estimates of $\hat{N}_X / \hat{N}_{Auto}$. Histograms show the distributions of $\hat{N}_X / \hat{N}_{Auto}$ estimates when the true N_X / N_{Auto} ratio is 0.75 estimated using unphased genotypes matching recombination rates (cM/Mb) to the high recombination dataset under the Schaffner demographic model. The gray bars are from simulations without any gene conversion on the X or the autosomes. The black bars are from simulated datasets which include gene conversion on the X chromosome but not on the autosomes (see Discussion).

Because it is likely that gene conversion occurs on both the X chromosome and the autosomes, these simulations are meant to be illustrative of the effect of differences in gene conversion on the X chromosome versus the autosomes on estimates of $\hat{N}_X / \hat{N}_{Auto}$, rather than being quantitative predictions about the amount of gene conversion on different compartments of the genome. It is apparent that gene conversion can lead to in increase in estimates of ρ using LDhat, and that differences

in the amount of gene conversion on the X chromosome versus the autosomes can account for our estimates of $\hat{N}_X / \hat{N}_{Auto}$ being higher than expected under models assuming no gene conversion (Figure 2.8). If this explanation is true, our elevated estimates of $\hat{N}_X / \hat{N}_{Auto}$ do not provide any meaningful information about sex-biased demography and cannot be taken as supporting either Hammer *et al.* (2008) or Keinan *et al.* (2009; see below).

There is some indirect evidence that there could be more gene conversion on the X chromosome than on the autosomes. First, there is an enrichment of inverted repeats on the X chromosome relative to the autosomes (Warburton *et al.* 2004). These repeats could lead to an increase in DNA mis-matches during meiosis which can in turn lead to gene conversion events during mis-match repair. Interestingly, the arms of the inverted repeats on the X chromosome are highly similar to each other even though the repeats themselves arose before the human-gorilla split and are quite old (Warburton *et al.* 2004). A similar pattern has been observed on the human Y chromosome, which contains a high fraction of palindromes where the two different arms of a given palindrome have near identical sequence (Skaletsky *et al.* 2003). Importantly, the Y chromosome palindromes are also quite old, suggesting that their integrity has been somehow maintained despite the absence of recombination (Rozen *et al.* 2003). Rozen *et al.* (2003) have demonstrated that the inter-arm identity of the Y chromosome palindromes is being maintained by inter-arm gene conversion. By analogy, a similar mechanism could be operating on the inverted repeats on the X chromosome. As a further similarity between the X and Y chromosome, both the Y chromosome palindromes and the X chromosome inverted repeats contain a number of tests expressed genes. Additional evidence for increased gene conversion on the X chromosome relative to the autosomes comes from reports of human X-Y gene conversion events (Rosser *et al.* 2009) as well as gene conversion among the color

vision genes on the X chromosome (Zhou and Li 1996). In principle, it should be possible to estimate levels of gene conversion on the X chromosome and the autosomes, however, previous attempts at estimating gene conversion from genetic variation data have found it difficult to obtain reliable results (Ptak *et al.* 2004). As such, this analysis may be better suited for the genome-wide resequencing datasets from a large number of individuals that will soon become available. Such an analysis should provide more direct evidence of the role of gene conversion on the X chromosome versus the autosomes.

Increased levels of gene conversion on the X chromosome relative to the autosomes may also be able to reconcile the previous studies of comparing patterns of variation on the X chromosome vs. the autosomes with the present study. We will discuss how each analysis may be affected by increased levels of gene conversion on the X relative to the autosomes. Our results based on LD patterns are likely to be affected in manner discussed above. Keinan *et al.* (2009) compared levels of population differentiation on the X chromosome and the autosomes as well as the frequency spectrum of SNPs ascertained within a single individual. Since these analyses all use relatively common SNPs, it is likely that they are not substantially affected by gene conversion, and could be reflective of demographic history. Hammer *et al.* (2008) compared the number of SNPs on the X vs. the autosomes in regions of high recombination rate. Evidence suggests that recombination in humans may be mutagenic (Lercher and Hurst 2002; Hellmann *et al.* 2003; Hellmann *et al.* 2005; Spencer *et al.* 2006) and that since recombination patterns differ between humans and chimps (Ptak *et al.* 2005; Winckler *et al.* 2005b), the mutation rate at a particular region may have changed over evolutionary time with the change in recombination patterns. This was the explanation offered by Hellmann *et al.* (2005) to explain why human diversity showed a better correlation with human recombination rates than

human-chimp divergence showed with human recombination rates. The effect of this phenomenon on nucleotide diversity is that nucleotide diversity will appear to be too high relative to what is expected based upon normalization by an outgroup. Combining this phenomenon with more gene conversion on the X chromosome than the autosomes could generate the pattern seen by Hammer *et al.* (2008). Hammer *et al.* (2008) selected regions on the X chromosome and the autosomes that had similar recombination rates based upon the deCODE genetic map (Hammer *et al.* 2008). Thus, all else being equal the excess number of SNPs relative to what is expected based on the outgroup normalization would be the same on the X and the autosomes. But, if gene conversion is also mutagenic and the X chromosome has an excess of gene conversion events, this could lead to an even higher mutation rate on the X chromosome that is not being taken into account by the outgroup normalization than on the autosomes. If true, this would lead to the pattern seen by Hammer *et al.* (2008), without including natural selection or an excess female effective population size. The nucleotide diversity analysis by Keinan *et al.* (2009) may be less affected by this phenomenon since they considered the entire chromosome, not just highly recombining regions.

While the above scenario offers a parsimonious reconciliation of the three studies, we cannot exclude the possibility that natural selection may have played an important role in affecting patterns of nucleotide diversity, the frequency spectrum, and population differentiation on the X chromosome and the autosomes. Further studies of large resequencing datasets from multiple human populations should provide additional evidence to resolve these questions.

Conclusion

We have shown that $\hat{N}_X / \hat{N}_{Auto} > 0.75$ when estimated from patterns of LD. If our results reflect only demographic history, then they are in full agreement with the

study of Hammer *et al.* (2008). If, however, elevated levels of gene conversion on the X chromosome relative to the autosomes contribute to our observations, then our results may also be consistent with those of Keinan *et al.* (2009).

CHAPTER 3

THE EFFET OF RECENT ADMXITURE ON INFERENCE OF ANCIENT POPULATION HISTORY[3]

3.1 Abstract

Despite the widespread study of genetic variation in admixed populations, such as African Americans, there has not been a systematic evaluation of the effects of recent admixture on patterns of polymorphism or inferences about population demography. These issues are particularly relevant because estimates of the timing and magnitude of population growth in Africa have differed among previous studies. It is unclear whether these differences can be explained by the fact that some studies included African American individuals while others sampled individuals from within Africa. Here we use simulations and human single nucleotide polymorphism (SNP) data collected through direct resequencing and genotyping to investigate these issues. We find that when estimating the current population size and magnitude of recent growth in an ancestral population using the site frequency spectrum (SFS), it is possible to obtain reasonably accurate estimates of the parameters based on samples drawn from the admixed population under certain conditions. We also show that methods for demographic inference that use haplotype patterns are more sensitive to recent admixture than are methods based on the SFS. The analysis of human genetic variation data from the Yoruba people of Ibadan, Nigeria (YRI) and African Americans supports the predictions from the simulations. Our results have important implications for the evaluation of previous population genetic studies that have

[3] Lohmueller, K.E., C.D. Bustamante, and A.G. Clark, Submitted.

considered African American individuals as a proxy for individuals from West Africa as well as for future population genetic studies of additional admixed populations.

3.2 Introduction

Studies of archeological and genetic data show that that anatomically modern humans originated in Africa and more recently left Africa to populate the rest of the world (Tishkoff and Williams 2002; Barbujani and Goldstein 2004; Garrigan and Hammer 2006; Reed and Tishkoff 2006; Campbell and Tishkoff 2008; Jakobsson *et al.* 2008; Li *et al.* 2008). Given the central role Africa has played in the origins of diverse human populations, understanding patterns of genetic variation within Africa and the demographic history of its populations is important for understanding the demographic history of global human populations. The availability of large-scale SNP datasets coupled with recent advances in statistical methodology for inferring population genetic models provides a powerful means of accomplishing these goals (Keinan *et al.* 2007; Boyko *et al.* 2008; Lohmueller *et al.* 2009; Nielsen *et al.* 2009). Understanding African demographic history is also important because researchers have used the demographic models inferred from genome-wide SNP data as a null model from which to compare regions of the genome. Regions that have patterns of variation inconsistent with the null model are often proposed as candidates to have undergone natural selection (Nielsen *et al.* 2009). Such approaches have identified numerous new potential candidate targets of selection in both African and non-African populations (Akey *et al.* 2004; Carlson *et al.* 2005; Stajich and Hahn 2005; Kelley *et al.* 2006; Voight *et al.* 2006; Kimura *et al.* 2007; Nielsen *et al.* 2007; Sabeti *et al.* 2007; Tang *et al.* 2007; Williamson *et al.* 2007; Nielsen *et al.* 2009; Pickrell *et al.* 2009).

It is important to realize that studies of African demographic history using genetic data have come to qualitatively different conclusions regarding important

parameters. Some recent studies have found evidence for ancient (>100,000 years ago) 2-4 fold growth in African populations (Adams and Hudson 2004; Marth *et al.* 2004; Keinan *et al.* 2007; Boyko *et al.* 2008). Other studies have found evidence for either very recent growth (Pluzhnikov *et al.* 2002; Akey *et al.* 2004; Voight *et al.* 2005; Cox *et al.* 2009; Wall *et al.* 2009), or no evidence of population growth at all (Pluzhnikov *et al.* 2002; Voight *et al.* 2005). It is unclear why these studies have found such different parameter estimates. However, these studies all differ from each other in the amount of data considered, the types of data used (e.g. SNP genotypes vs. full resequencing), the genomic regions studied (e.g. non-coding vs. coding SNPs) and the types of demographic models considered (e.g. including migration vs. not including migration post-separation of African and non-African population).

Another important way in which studies of African demographic history differ from each other is in the populations sampled. Some studies have focused on genetic data from individuals sampled from within Africa, while other studies included American individuals with African ancestry. While there is no clear correspondence between those studies which sampled native African individuals (as opposed to African Americans) and particular growth scenarios, it is clear from previous studies that African American populations do differ from African populations in their recent demographic history. In particular, genetic studies suggest that there is wide variation in the degree of European admixture in most African American individuals in the U.S., and that they have, on average, ~80% African ancestry and 20% European ancestry (Parra *et al.* 1998; Pfaff *et al.* 2001; Falush *et al.* 2003; Patterson *et al.* 2004; Tian *et al.* 2006; Lind *et al.* 2007; Reiner *et al.* 2007; Price *et al.* 2009; Bryc *et al.* 2010). Furthermore, both historical records and genetic evidence suggest that the admixture process began quite recently, within the last 20 generations (Pfaff *et al.* 2001; Patterson *et al.* 2004; Seldin *et al.* 2004; Tian *et al.* 2006). Recent population

admixture can alter patterns of genetic variation in a discernable and predictable way. For example, recently admixed populations will exhibit correlation in allele frequencies (i.e., linkage disequilibrium) among markers that differ in frequency between the parental populations. This so-called admixture linkage disequilibrium (LD; Chakraborty and Weiss 1988) can extend over long physical distances (Lautenberger *et al.* 2000) and decays exponentially with the time since the admixture process began (i.e., recently admixed populations typically exhibit LD over a longer physical distance than anciently admixed populations). In fact, commonly used unsupervised algorithms for identifying population structure and admixture (such as the Bayesian clustering algorithm STRUCTURE; see Falush *et al.* 2003) exploit admixture LD to assign blocks of the genomes of admixed individuals to the ancestral populations (Falush *et al.* 2003; Patterson *et al.* 2004; Tang *et al.* 2006; Sankararaman *et al.* 2008a; Sankararaman *et al.* 2008b; Price *et al.* 2009; Bryc *et al.* 2010).

While it is clear that African American populations have a different recent demographic history than do African populations from within Africa, and that admixture tracts can be identified in admixed individuals, the effect that admixture has had on other patterns of genetic variation remains unclear. For example, Xu *et al.* (2007) found similar LD decay patterns when comparing African American and African populations. It is also unclear whether the recent admixture impacts our ability to reconstruct ancient demographic events (such as expansions that predate the spread of humans out of Africa) from whole genome SNP data. Most studies of demographic history have summarized the genome-wide SNP data by allele frequency or haplotype summary statistics. If these summary statistics are not sensitive to the recent European admixture, then the African American samples may yield estimates of demographic parameters that are close to the true demographic parameters for the ancestral, un-sampled, African populations. This would suggest that the differences in

growth parameter estimates obtained from African populations cannot be explained by certain studies sampling African American individuals and others sampling African individuals from within Africa. However, if these statistics are sensitive to recent admixture, then they may give biased estimates of growth parameters.

Here, we examine the effect of recent admixture on the estimation of population demography. In particular, we estimate growth parameters from simulated datasets using SNP frequencies as well as a recently developed haplotype summary statistic (Lohmueller *et al.* 2009). We compare the demographic parameter estimates made from the admixed and non-admixed populations and find that some parameter estimates are qualitatively similar between the two populations when inferred using allele frequencies. Inferences of growth using haplotype-based approaches appear to be more sensitive to recent admixture than inferences based on SNP frequencies. We discuss implications that our results have for interpreting studies of demography in admixed populations.

3.3 Methods

Demographic model for simulations

For generating simulated data, we used a demographic model that qualitatively approximates the history of African, European, and African American human populations. We chose to focus on African American demography as 1) African American populations are a significant component of the U.S. population (~12%, U.S. Census 2000 Summary File 1, http://factfinder.census.gov/servlet/) and are, therefore, heavily studied by population and medical geneticists in the U.S., 2) There is considerable understanding of the historical context surrounding the recent demographic history of African Americans including the trans-Atlantic slave trade, early American history, and history of African-American migrations within the U.S., and 3) the admixture process in other human populations is likely to be more complex.

Figure 3.1 shows an illustration of the demographic model considered. Essentially, an ancestral population of size N_B split t_{split} generations ago to form an African population (Pop A) and a European population (Pop E).

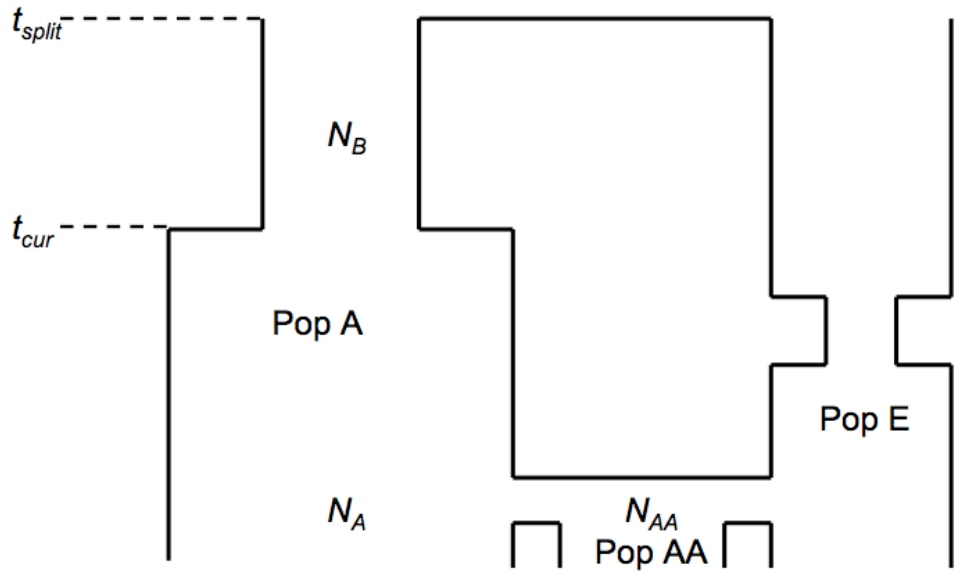


Figure 3.1: Demographic model for African (Pop A), African American (Pop AA), and European (Pop E) populations used to simulate test datasets. For all simulations conducted here, t_{split} = 4000 generations, N_a = 20,000, N_B = 10,000 and Pop E followed a bottleneck model from Lohmueller *et al.* (2009), except the ancestral size, which was set to 10,000. See text for a further description of the parameters.

The African population expanded from its ancestral size (N_B) to its current size (N_A) t_{cur} generations ago. The European population underwent a bottleneck (using parameters similar to those inferred by Lohmueller *et al.* 2009). Note, we assume no gene flow between the African and European populations after the split even though some studies have found evidence for migration between African and European populations (Schaffner *et al.* 2005; Nielsen *et al.* 2009; Wall *et al.* 2009; Gutenkunst *et al.* 2009). We chose not to include such migration in our models so that our

assessments of the effects of recent admixture would not be confounded by other sources of gene flow. Twenty generations ago, the African American population (Pop AA) is formed and has current size N_{AA} (Pfaff *et al.* 2001; Patterson *et al.* 2004; Tian *et al.* 2006). We assume 80% of the ancestry of Pop AA comes from Pop A, with the remainer coming from Pop E (Pfaff *et al.* 2001; Patterson *et al.* 2004; Tian *et al.* 2006). Since it is unknown whether there was a founder effect in forming the African American population, we allowed N_{AA} to vary. All simulations assume an infinite sites mutation model, random mating within each population, and no natural selection.

Inference on simulated data using the site frequency spectrum (SFS)

A useful summary of SNP data that potentially contains information regarding the magnitude of recent population growth is the site frequency spectrum (SFS; Fu 1995; Griffiths and Tavaré 1998; Nielsen 2000; Williamson *et al.* 2005). Mathematically, the SFS is defined for a set of n sequenced chromosomes across S variable sites as the random vector $(X_1, X_2, \ldots X_{n-1})$ where X_i represents the number of sites (i.e., SNPs) where the n chromosomes are portioned into exactly i copies of the derived allele and n-i copies of the ancestral allele. For example, X_1 is the number of singletons SNPs in the data, X_2 is the number SNPs where exactly two chromosomes carry the derived SNP, and so on. Note that the sum of the entries in the SFS equals the total number of SNPs in the dataset. Informally, one can think of the SFS as a histogram consisting of the number of SNPs at different frequencies in the sample where frequencies are binned at $1/n$ intervals. To determine the accuracy of parameter estimates for N_A, t_{cur}, and N_B/N_A when using the SFS obtained from Pop AA, for each combination of demographic parameters, we simulated 500 datasets, each consisting of 10,000 unlinked 1kb regions in $n = 24$ chromosomes from each population. The size of each dataset was meant to mimic the scope of resequencing datasets currently in use, such as the Celera Genomics SNP dataset (Bustamante *et al.* 2005; Lohmueller

et al. 2008). We assume a per-nucleotide mutation rate $\mu = 10^{-8}$ and a per-nucleotide recombination rate, $r = 10^{-8}$. For each dataset, we calculate the SFS for both Pop A and Pop AA which are then used for inference.

To estimate the MLEs for the three growth parameters, we use a Poisson likelihood function (see Nielsen 2000; Williamson *et al.* 2005; Boyko *et al.* 2008) for details). Briefly, the observed number of SNPs in each bin, X_i, of the SFS is treated as a Poisson random variable:

$$\Pr(X_i = k \mid \Theta) = e^{-\lambda_i} \lambda_i^k / k! \qquad (3.1)$$

where the rate parameter of the Poisson distribution λ_i is the expected number of SNPs in the particular bin of the SFS based on the set $\Theta = \{\theta, \upsilon, \tau\}$ of mutation rate ($\theta = 4N_A\mu$) and growth ($\upsilon = N_B / N_A$; $\tau = t_{cur} / 2N_A$) parameters. This expectation has the form where the mutation rate acts as a scaling factor for the whole region, so we can rewrite as the product as

$$\lambda_i = \mathrm{E}(X_i \mid \Theta) = \theta F(i \mid \upsilon, \tau) \qquad (3.2)$$

where $F(i \mid \upsilon, \tau)$ is proportional to the number of SNPs at frequency *i*/*n* in the sample, and can be found either by coalescent simulations (Nielsen 2000) or via diffusion based approximations (Williamson *et al.* 2005). All bins of the SFS are treated independently and the final log-likelihood for a given set of growth parameters is the sum of the Poisson log-likelihoods for each bin of the SFS as given below:

$$l(\Theta \mid x) = -\theta \sum_{i=1}^{n-1} F(i \mid \tau, \upsilon) + x_i \log\left(\theta F(i \mid \tau, \upsilon)\right) \qquad (3.3)$$

This is a reasonable approximation to the true log-likelihood and holds when there is ample recombination among SNPs. Since the expected value of the SFS entries are not impacted by recombination, when applied to linked data, the above inference scheme can be thought of as a composite likelihood (Zhu and Bustamante 2005; Boyko *et al.* 2008).

We use the program prfreq (Boyko *et al.* 2008) to find the expected SFS for a given set of growth parameters, using the Poisson rather than the multinomial implementation. The multinomial would maximize the function above for the θ term and allow for inference solely on the "curvature" of the SFS rather than on the actual counts observed. Here we set $\mu = 0.1$, or the true value used in the simulations to generate the data (10^{-8} per bp x 10^3 bp per region x 10^4 regions), and in doing so, estimate the unscaled parameters. To optimize the likelihood function, we found the expected SFS for each parameter combination on a 3-dimensional grid (N_A, t_{cur}, and N_B/N_A) of parameter values. It should be noted that the grids used are coarser than those used on real datasets (see below) and that fixing of μ to a particular value does not alter the coverage properties for scaled parameters due to the invariance principle of maximum likelihood inference (Pawitan 2001).

For each dataset, we calculated approximate 95% confidence intervals using the log-likelihood curves without any smoothing or interpolation. Single-parameter CIs included grid points within 1.92 log-likelihood of the MLE using the profile-likelihood curves. This corresponds to defining the acceptance region for a one-dimensional likelihood-ratio test for the parameter being evaluated under the null hypothesis that the observed value is the true value and that the test statistic (2*log-likelihood differences from the MLE) follows a chi-square with one degree of freedom (i.e., critical value of 3.84). The three-dimensional CI is similarly defined as the convex hull which includes all grid points within 3.9 log-likelihood units of the MLE in three-dimensional space and corresponds to the critical value of 7.81 for a likelihood-ratio test with asymptotic distribution of a chi-square with three degrees of freedom. To obtain accurate CIs, an accurate estimate of the likelihood surface around the MLEs is required. This is often done using a grid-search technique in an iterative manner, increasing the density of the grid points near the current MLEs. Since we

were analyzing thousands of simulated datasets, it was impractical to do such an iterative grid search on every dataset. Therefore, the CIs in the simulated datasets are based on using coarse grids and should be regarded as approximate. Nevertheless, the coverage properties of these intervals serve as a useful heuristic.

Inference on simulated data using the haplotype-count number (HCN) statistic

Lohmueller *et al.* (2009) recently suggested a new summary statistic for genome-wide SNP data based on haplotype patterns. Their statistic, termed the HCN statistic, is a two-dimensional histogram containing the joint distribution of the number of haplotypes and count of the most common haplotype in windows across the genome. To determine whether we could accurately estimate N_A, t_{cur}, and N_B/N_A when using the HCN obtained from Pop AA, for each combination of demographic parameters, we simulated 100 datasets, each consisting of 7,000 unlinked 250 kb regions in $n = 46$ chromosomes from each population. Importantly, as in Lohmueller *et al.* (2009), we only used a random subset of 20 SNPs from each simulated window with minor allele frequency (MAF) >10%. The size of these datasets is meant to mimic the large-scale genotyping surveys currently in use, such as that of Perlegen Sciences, where only a subset of SNPs in the population have been discovered and genotyped. We assume a per-nucleotide mutation rate $\mu = 10^{-8}$ and a per-nucleotide recombination rate, $r = 10^{-8}$. For each dataset, we calculated the HCN statistics for Pop A and Pop AA which are then used for inference.

To find the MLEs of the three demographic parameters from each dataset via the HCN statistic, we use the approach of Lohmueller *et al.* (2009). Briefly, we use a multinomial approximate likelihood function where the observed number of regions with i haplotypes and where the most common haplotype is at count j comes from a multinomial distribution whose parameters are determined by the recombination rate and demographic parameters. We use coalescent simulations to find the parameters

for the multinomial distribution for a given set of demographic parameters and recombination rates. Importantly, we fixed the recombination rate in these simulations equal to the true value used to generate the test datasets. The likelihood function is optimized using a grid search. We found approximate 95% CIs for the HCN based demographic estimates using the same procedure as described above for the SFS-based CIs. Again, since the CIs in the simulated datasets are based on using coarse grids, the coverage properties of these intervals should serve as a useful heuristic.

Analysis of NIEHS data

We fitted a growth model to the SFS of the Yoruba (YRI) and African American (AA) samples from the NIEHS data (Livingston *et al.* 2004). The YRI sample contained $n = 12$ individuals from Ibadan, Nigeria and the AA sample contained 15 individuals sampled from the U.S. We only used noncoding SNPs and excluded SNPs that had genotypes for fewer than 10 individuals in at least one of the two populations. Since some SNPs did not have a genotype at every individual, we used the hypergeometric distribution to find the expected SFS for a sample size of 20 chromosomes (Nielsen *et al.* 2004). In total our analysis included 13,588.4 SNPs in the AA population and 13,487.9 SNPs in the YRI population after projection to a sample size of 20 chromosomes. The projection and subsequent analyses were done on the AA and YRI samples separately. For the analysis of the NIEHS data, we used the folded SFS. The folded SFS tabulates the frequency of the minor allele, rather than the derived allele. The folded SFS still contains substantial information regarding demography without having to accurately infer the ancestral/derived states of SNPs (Adams and Hudson 2004).

We estimated the growth parameters for the YRI and AA datasets using the prfreq program (Boyko *et al.* 2008). Like we did for the analysis of the simulated datasets, we used the Poisson likelihood function. This procedure requires an accurate

estimate of the per-nucleotide mutation rate, μ. To estimate μ, we used the level of human-chimp divergence at the region sequenced and the relationship $K = 2T\mu$, where K is the number of human-chimp differences (per nucleotide) and T is the human-chimp divergence time in units of generations. There were 51,770 differences in the 4,644,887 nucleotides sequenced in the builds of the human (hg18) and chimpanzee (pantro2) genomes, giving $K = 0.01114559$ per nucleotide. Then, assuming a human-chimp divergence time of 6 million years, and 25 years/generation,

$$\mu = \frac{0.01114559 \times 25}{2 \times 6 \times 10^6} = 2.32 \times 10^{-8} \qquad (3.4)$$

per nucleotide/generation. Since before the hypergeometric projection of the SNP frequencies, 7.83% of SNPs were excluded because they contained genotypes for fewer than 20 chromosomes in either one population or both populations, we decreased the total number of nucleotides sequenced by the same amount. This led to 4,281,191 total nucleotides that were used for analysis. We again used a grid search to optimize the likelihood function.

Analysis of the Perlegen data

We fitted a growth model to the AA sample analyzed by Perlegen Sciences (Hinds *et al.* 2005) using the HCN statistic. We chose to use the Perlegen data rather than other datasets, such as HapMap, because Perlegen genotyped all SNPs that they discovered, without regard to LD status, making subsequent analyses simpler (Lohmueller *et al.* 2009) and relatively free of the ascertainment biases in HapMap.

We divided the genome into non-overlapping 0.25 cM windows and selected 20 SNPs from each window to construct the HCN statistic. Note, we only selected SNPs with MAF >10% in both the AA and CEU datasets. Additionally, SNPs that were discovered using <8 chromosomes were not included in the analysis. In total, the

HCN statistic contained 8174 windows. As in Lohmueller *et al.* (2009), we used Clark's phasing algorithm (Clark 1990) to infer haplotype phases of the SNP data.

In the coalescent simulations used to generate the expected HCN statistic for a given demographic model, we also phased the simulated data using Clark's phasing algorithm and drew the recombination rate for each simulated region for a gamma distribution to allow for errors in the estimated genetic map (Lohmueller *et al.* 2009). Finally, we used the Schaffner recombination hotspot model (Schaffner *et al.* 2005) as implemented in Lohmueller *et al.* (2009).

3.4 Results

Effect of admixture on patterns of polymorphism

Figure 3.2 shows the expected SFS for Pop A with 2-fold growth as well as Pop AA under two different values of N_{AA}. First, we note there is an excess of low-frequency SNPs in Pop A compared to the neutral prediction. This result is expected since an excess of low-frequency SNPs is a signature of the population expansion (Tajima 1989a; Slatkin and Hudson 1991). For Pop AA, when $N_{AA} = N_A$, there is an even more pronounced excess of singleton and doubleton SNPs over what is seen in Pop A. However, the remaining bins of the SFS are similar for Pop A and Pop AA. When $N_{AA} = 0.1N_A$, we observe a decrease in the number of singleton SNPs compared to when $N_{AA} = N_A$. However, the number of singleton SNPs with $N_{AA} = 0.1N_A$ is still slightly greater than that for Pop A alone. We also examined the SFS when Pop AA was formed 7 generations ago (Price *et al.* 2009) instead of 20 generations ago, corresponding to more recent admixture (Figure 3.3). When $N_{AA} = N_A$, the SFS for the two different admixture times are identical. When $N_{AA} = 0.1N_A$, more recent admixture results in a slight increase in the number of singletons, presumably since the more recent founding of Pop AA results in less drift in Pop AA.

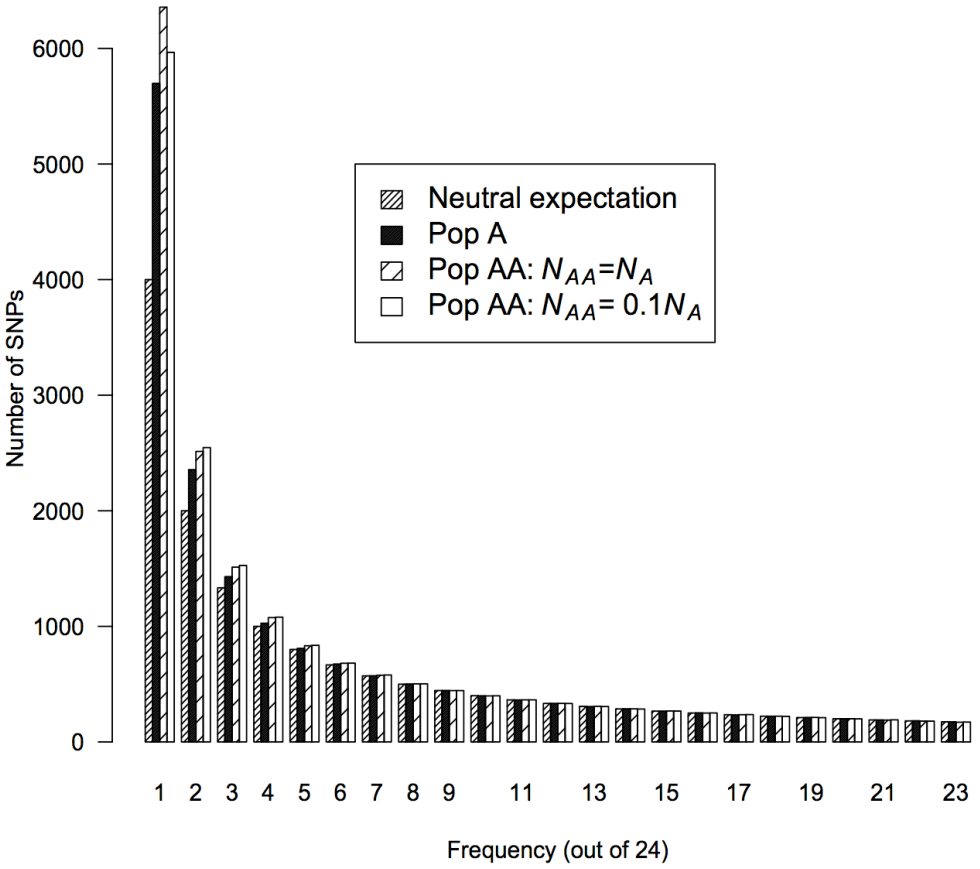


Figure 3.2: Expected SFS in a sample size of 24 chromosomes for Pop A and Pop AA under population growth. Note the excess of low frequency SNPs relative to the neutral prediction (Fu 1995) in all populations as well as the more pronounced excess of low frequency SNPs in Pop AA relative to Pop A. N_A = 20,000; N_B/N_A = 0.5; t_{cur} = 2400 generations.

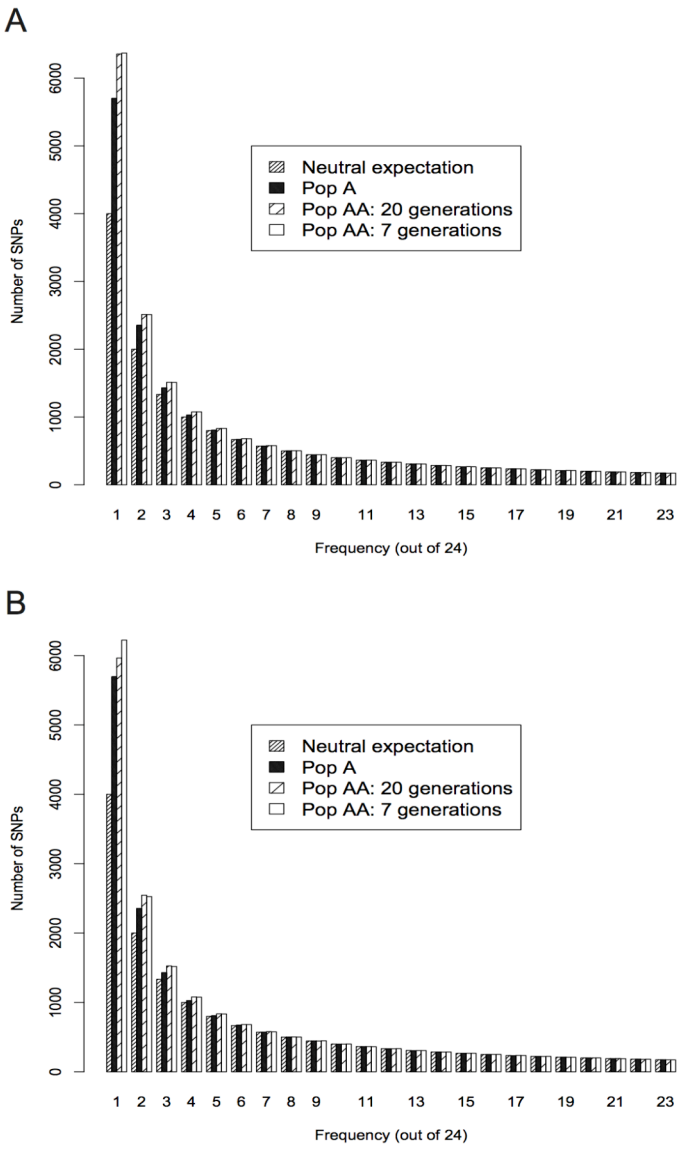


Figure 3.3: Expected SFS in a sample size of 24 chromosomes for Pop A and Pop AA under population growth when admixture occurs 20 or 7 generations ago. Note in both cases, the SFS when admixture occurs 20 generations ago is similar to that when admixture occurs 7 generations ago. A). $N_{AA} = N_A$ and B). $N_{AA} = 0.1N_A$. Here there is a slight excess of singletons when admixture occurs 7 generations ago as opposed to 20 generations ago, since there is less drift in Pop AA with the more recent founding. However, the number of singletons in Pop AA when $N_{AA} = 0.1N_A$ is still less than that when $N_{AA} = N_A$. $N_A = 20{,}000$; $N_B/N_A = 0.5$; $t_{cur} = 2400$ generations.

However, the number of singletons here is still lower than that seen in Pop AA when $N_{AA} = N_A$. Thus, for the parameter combinations investigated here, the site frequency

spectra for Pop A and Pop AA all appear to have an excess of low-frequency SNPs, with the excess being more pronounced in Pop AA.

We also investigated the effect of recent admixture on the HCN statistic. Figure 3.4 shows the HCN statistics for simulated data under the three models described above (Pop A; Pop AA: $N_{AA} = N_A$; Pop AA: $N_{AA} = 0.1N_A$). For Pop AA when $N_{AA} = N_A$, there is an excess of regions that have fewer haplotypes as well as a shift toward more regions having the most common haplotype at higher frequency relative to Pop A. This pattern likely stems from the fact that some individuals in the admixed population have haplotypes which recently came from Pop E. Since Pop E underwent a bottleneck, it contains less haplotype diversity than Pop A. Thus, in the HCN for the admixed population, the Pop E haplotypes create a shift toward more regions with fewer haplotypes and where the most common haplotype is at higher frequency. When $N_{AA} = 0.1N_A$, the differences in the HCN statistic between Pop AA and Pop A become even more pronounced. Here we observe more regions with fewer haplotypes and where the most common haplotype is at higher frequency relative to the scenario for Pop AA when the ancestral and admixed populations are identical in size ($N_{AA} = N_A$). This pattern is likely due to the loss of haplotypes when the size of Pop AA decreases. We also examined the HCN statistic when Pop AA was formed 7 generations ago (Price *et al.* 2009) instead of 20 generations ago, corresponding to more recent admixture (Figure 3.5). When $N_{AA} = N_A$, the HCNs for the two different admixture times are nearly identical. When $N_{AA} = 0.1N_A$, more recent admixture results in less of a shift towards more regions with fewer haplotypes and where the most common haplotype is at higher frequency, presumably since the more recent founding of Pop AA results in less drift in Pop AA.

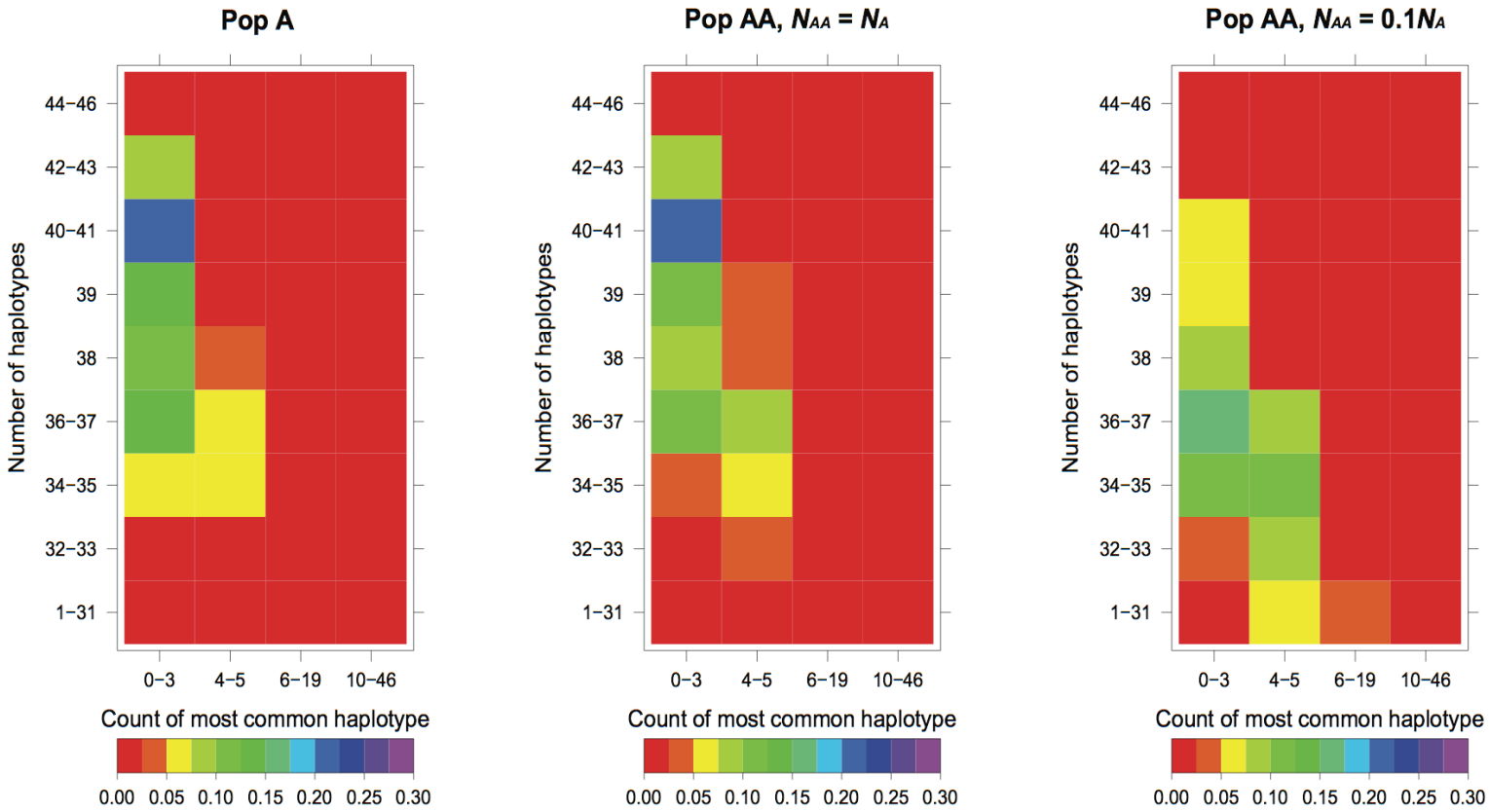


Figure 3.4: Expected HCN statistic for Pop A and Pop AA. Each cell in the matrix is colored according to the proportion of simulation replicates having the particular configuration of the number of haplotypes and count of the most common haplotype. For example, red cells contain <1% of regions and dark blue cells contain 20-25% of simulated regions. Note the excess of regions with fewer haplotypes and where the most common haplotype is at higher frequency in Pop AA relative to Pop A. This is most pronounced when $N_{AA} = 0.1N_A$. Note, the simulations assume $N_A = 20{,}000$; $N_B/N_A = 0.5$; $t_{cur} = 2400$ generations.

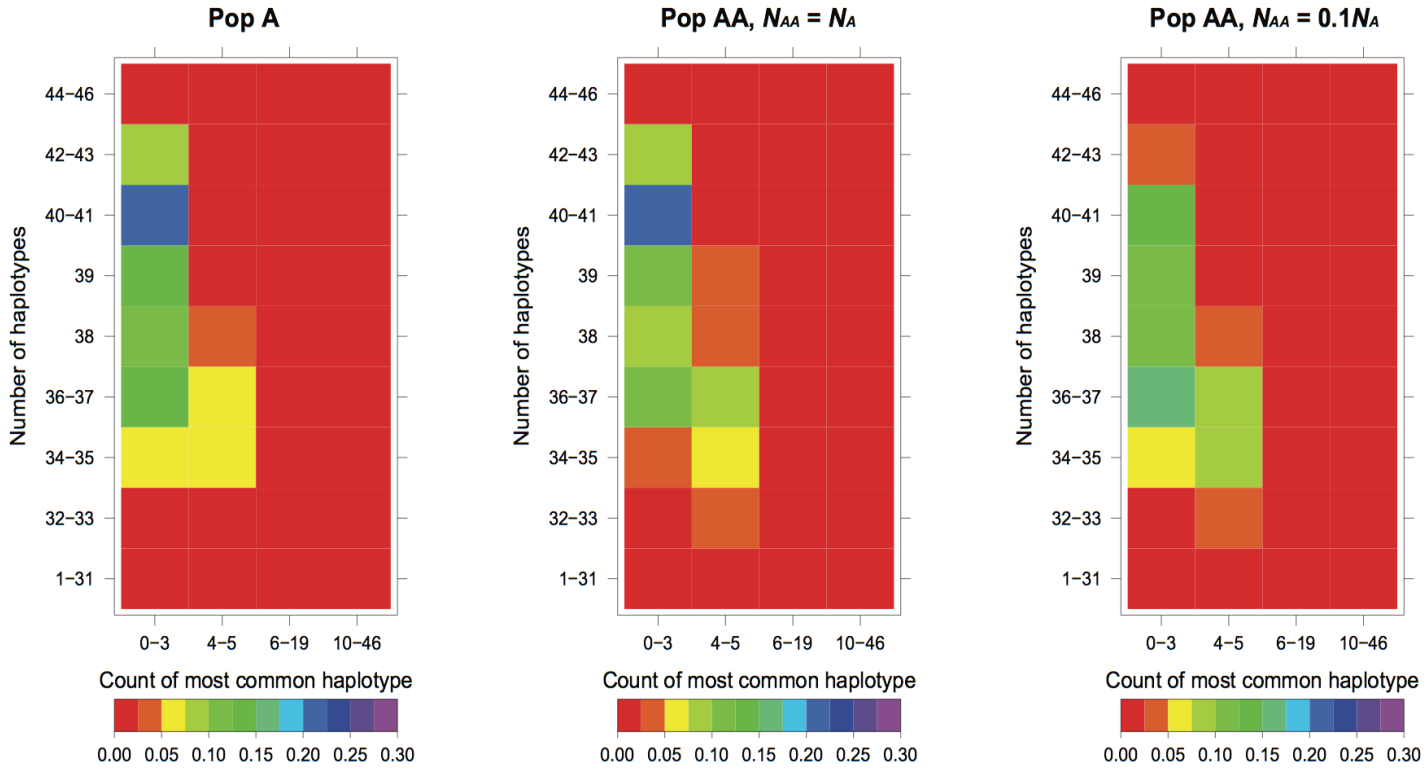


Figure 3.5: Expected HCN statistic for population growth (Pop A) and for growth with admixture (Pop AA) when admixture occurred 7 generations ago (instead of 20 generations). Each cell in the matrix is colored according to the proportion of simulation replicates having the particular configuration of the number of haplotypes and count of the most common haplotype. For example, red cells contain <1% of regions and dark blue cells contain 20-25% of simulated regions. Note the excess of regions with fewer haplotypes and where the most common haplotype is at higher frequency in Pop AA relative to Pop A. This is most pronounced when $N_{AA} = 0.1N_A$. Note, the simulations assume $N_A = 20{,}000$; $N_B/N_A = 0.5$; $t_{cur} = 2400$ generations.

However, the overall haplotype diversity is still lower than that seen in Pop AA when $N_{AA} = N_A$. In summary, the admixture process alters the HCN statistic in a manner that is heavily influenced by the current size of Pop AA and the relative duration of the founder effect forming Pop AA.

Inference of demography from simulated data

To determine whether the differences in the SFSs and HCNs from Pop A and Pop AA (Figures 3.2-3.5) are meaningful, we estimated the parameters for a population growth model using the SFSs and HCNs from Pop A and Pop AA (see Methods). The purpose of this analysis was to see if using data from Pop AA, we could accurately estimate the current population size of the Pop A (N_A), time of population growth in Pop A (t_{cur}) and the magnitude of population growth (N_B/N_A).

Figure 3.6A shows the distribution of MLEs inferred using the SFS for the three growth parameters when $t_{cur} = 2400$ generations. For Pop A, the MLEs for all three parameters are clustered at the true parameter values. For Pop AA, when $N_{AA} = N_A$, N_A is slightly over-estimated and N_B/N_A is under-estimated. The reverse pattern is seen when $N_{AA} = 0.1N_A$. Here, N_A is slightly underestimated, but N_B/N_A is slightly over-estimated. Importantly, in both cases, when using the SFS from Pop AA, all the estimates for N_A are within 10,000 of the true value and all the estimates for N_B/N_A are within 0.15 of the true value. The estimates of the timing since the instantaneous growth event (t_{cur}) present a different pattern. For both models of Pop AA, t_{cur} is severely over-estimated (see below).

Figure 3.6B shows the distribution of the MLEs inferred using the HCN method. Again, the MLEs from Pop A are clustered around the true parameter values. Unlike for the estimates made using the SFS, the MLEs from Pop AA are now quite far from the true parameter values. For example, when $N_{AA} = N_A$, N_A is severely over-estimated, and N_B/N_A is under-estimated. When $N_{AA} = 0.1N_A$, haplotype diversity is

lost, leading to the under-estimation of N_A when using individuals sampled from Pop AA. Interestingly, t_{cur} is now under-estimated in both cases, presumably due to the fact that the very recent admixture has affected the haplotype patterns.

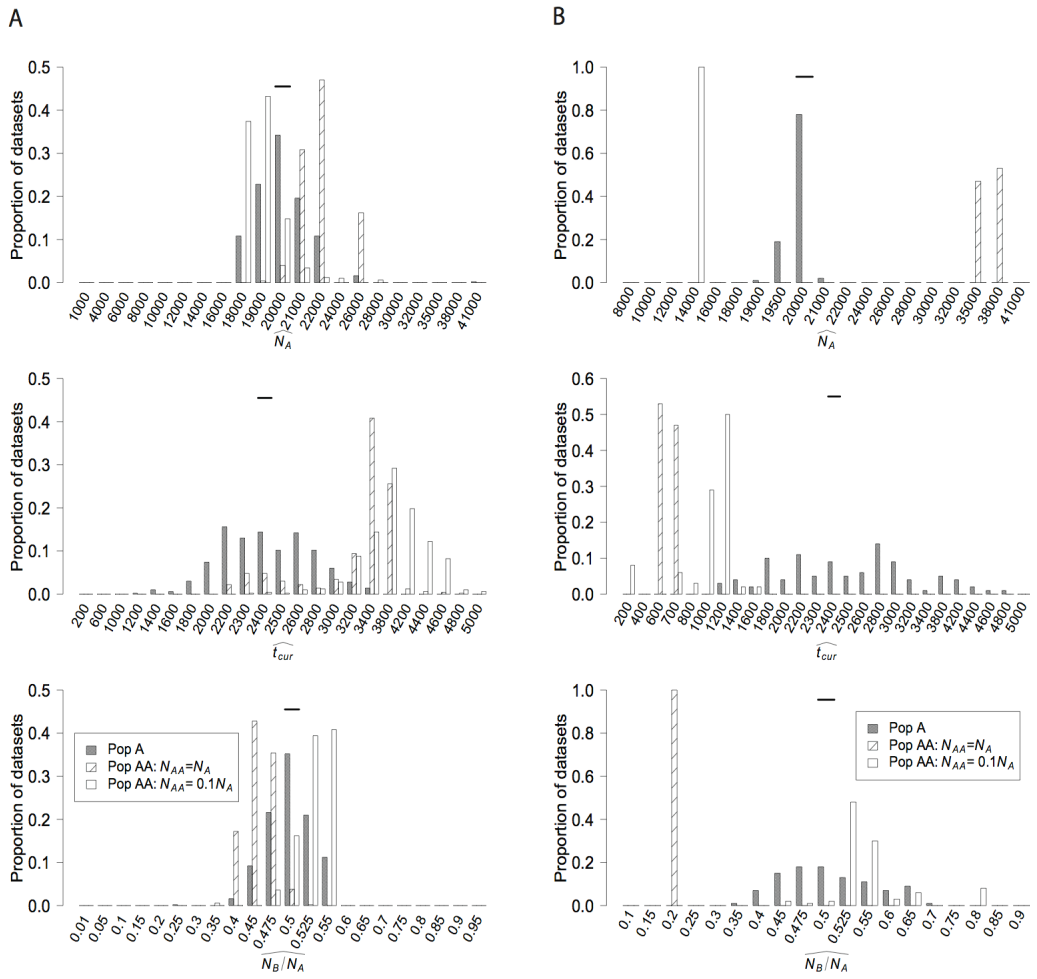


Figure 3.6: Distribution of MLEs for the three growth parameters inferred using A) the SFS and B) the HCN method (see text). Solid horizontal bars denote the true parameter values (N_A = 20,000; N_B/N_A = 0.5; t_{cur} = 2400 generations).

We also analyzed additional simulated datasets where t_{cur} = 4000 generations. Figure 3.7 shows the distribution of the MLEs inferred using the SFS (Figure 3.7A) and the HCN statistic (Figure 3.7B). Note that when estimating parameters using the

SFS, the estimates made from Pop AA again closely approximate the true parameter values for Pop A.

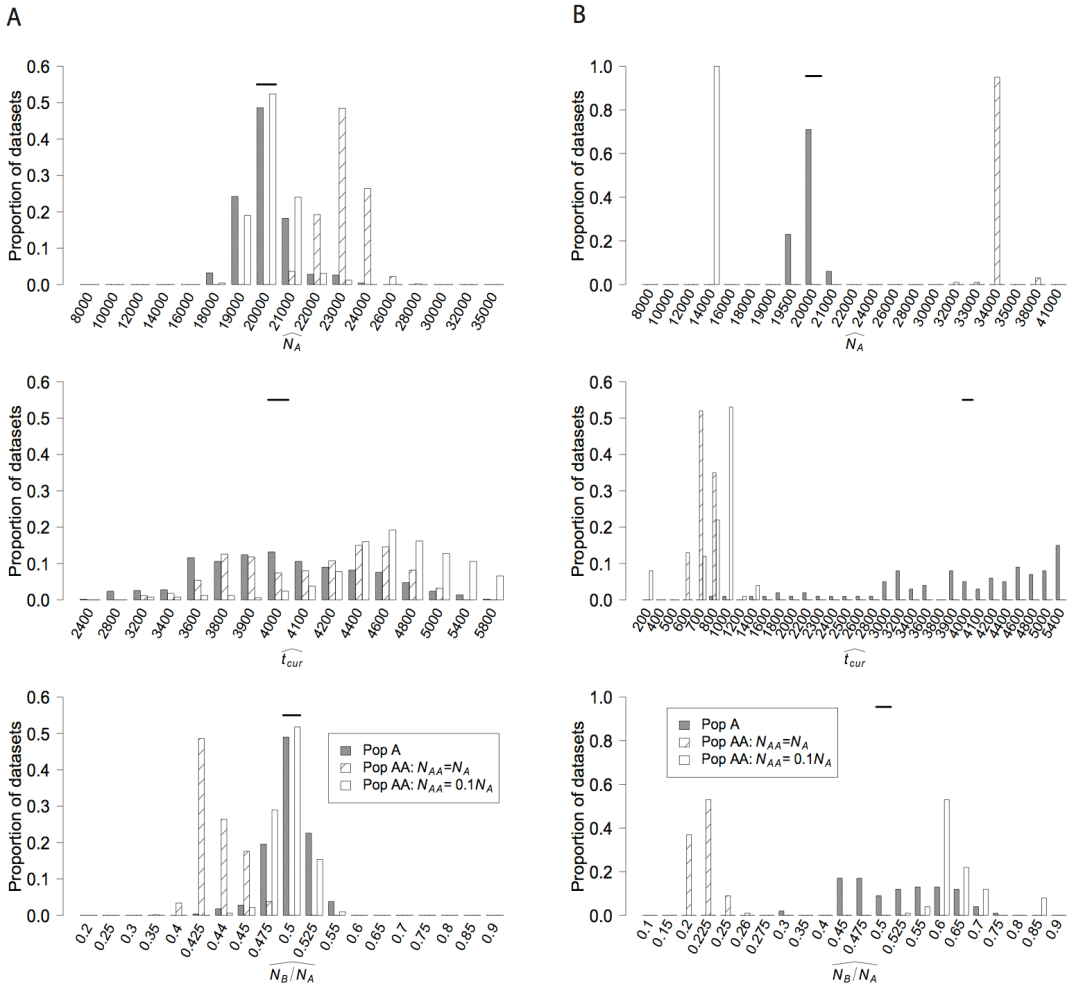


Figure 3.7: Distribution of MLEs for the three growth parameters inferred using A) the SFS and B) the HCN method (see text). Solid horizontal bars denote the true parameter values (N_A = 20,000; N_B/N_A = 0.5; t_{cur} = 4000 generations).

In particular, t_{cur} is not as severely over-estimated as compared to the case where t_{cur} = 2400 generations. This is especially noticeable when $N_{AA} = N_A$. Thus, part of the explanation for the over-estimate of t_{cur} using Pop AA when t_{cur} = 2400 generations is that the estimate was heavily influenced by the population "expansion" that occurred

at t_{split}, when the European and African populations split (Stadler *et al.* 2009). When $t_{cur} = t_{split}$, this over-estimate is less pronounced, although is still present when $N_{AA} = 0.1N_A$.

Figure 3.7B shows that the MLEs inferred using the HCN statistic on individuals from Pop AA are again very far from the true growth parameter values in Pop A. Interestingly, the MLEs of the growth parameters inferred from Pop AA when $t_{cur} = 4000$ are very similar to those inferred when $t_{cur} = 2400$ (compare Figure 3.6B to Figure 3.7B). This suggests that the admixture process has such a profound influence on the haplotype patterns that changes in the timing of growth in the parental population cannot be detected using individuals from Pop AA. The current size of Pop AA (N_{AA}), on the other hand, has a large impact on the haplotype patterns (compare Pop AA: $N_{AA} = N_A$ to Pop AA: $N_{AA} = 0.1N_A$ in Figure 3.7B). These results, taken together, indicate that recent admixture affects haplotype summary statistics more than it affects the SFS.

In addition to examining the distribution of the MLEs of the growth parameters, we also assessed coverage properties of approximate 95% confidence intervals (CIs) made from the profile likelihood curves. It should be noted that these CIs are likely to be anti-conservative (i.e. contain the true parameter values <95% of the time) since the asymptotic results assume that all the observations are independent. However, in practice, SNPs are not completely independent. Table 3.1 shows the percent of the time that the three-dimensional approximate CI contains all three growth parameter values or the one-dimensional CI from the profile likelihood curves contain the individual parameter values. We find that for inference using the SFS, the 95% CIs from Pop A contain the true parameter value roughly 95% of the time, although, as expected, the CIs are sometimes anti-conservative. The three-dimensional CIs made using the Pop AA SFSs rarely contain the true parameter

values. The picture for the single-parameter CIs is a bit more encouraging and variable. For example, when t_{cur} = 4000 and $N_{AA} = 0.1N_A$, the CI for N_B/N_A contains the true parameter value 93.8% of the time, which is slightly higher than the coverage

Table 3.1: Coverage properties of approximate 95% CIs for growth parameters estimated from Pop A and Pop AA using the SFS and HCN methods[a].

		Population[b]	Overall[c]	$\hat{N}_A$	$\frac{\hat{N}_B}{N_A}$	$\hat{t}_{cur}$
SFS	$t_{cur} = 2400$	Pop A	96%	93.80%	94.60%	93.00%
		Pop AA, $N_{AA} = N_A$	0%	37.20%	36.80%	29.20%
		Pop AA, $N_{AA} = 0.1N_A$	0%	70.40%	70.60%	6.60%
	$t_{cur} = 4000$	Pop A	95.40%	92.40%	92.20%	89.60%
		Pop AA, $N_{AA} = N_A$	0%	1.80%	1.60%	82.20%
		Pop AA, $N_{AA} = 0.1N_A$	25.20%	93.60%	93.80%	44.20%
HCN	$t_{cur} = 2400$	Pop A	100%	100%	100%	100%
		Pop AA, $N_{AA} = N_A$	0%	0%	0%	0%
		Pop AA, $N_{AA} = 0.1N_A$	0%	0%	63%	0%
	$t_{cur} = 4000$	Pop A	100%	100%	100%	99%
		Pop AA, $N_{AA} = N_A$	0%	0%	0%	0%
		Pop AA, $N_{AA} = 0.1N_A$	0%	0%	21%	0%

a. In all cases, the true parameters are N_A = 20,000 and N_B/N_A = 0.5.
b. The population used for inference. "Pop A" denotes the case where the non-admixed population is use to infer the growth parameters. "Pop AA, $N_{AA} = N_A$" denotes the case where samples from Pop AA are used to estimate growth parameters in Pop A, and the current size of Pop AA is equal to that of Pop A. "Pop AA, $N_{AA} = 0.1N_A$" denotes the case where samples from Pop AA are used to estimate growth parameters in Pop A, and the current size of Pop AA is 0.1 that of Pop A.
c. The three-dimensional 95% CI.

from Pop A (92.2%). Unfortunately, coverage is parameter-value dependent: when t_{cur} = 4000, and $N_{AA} = N_A$, the CI for N_B/N_A contains the true parameter value only 1.6% of

the time. The fact that the single-parameter CIs often contain the true parameter values, but the joint intervals do not suggests that many datasets accurately estimate one or two of the parameters, but not all three. For inference using the HCN statistic, we find that for Pop A, the 95% CIs are conservative, consistent with the observations of Lohmueller *et al.* (2009). However, the CIs made using data from Pop AA rarely contained the true parameter values from Pop A.

Frequently when researchers fit a demographic model to the observed SFS, they will also perform a goodness-of-fit (GOF) test to determine if the best-fitting model can explain observed SFS (see for example Adams and Hudson 2004; Caicedo *et al.* 2007; Boyko *et al.* 2008; Nielsen *et al.* 2009). Given that the simple growth model is the wrong model for Pop AA (the true model involves growth and admixture), we wanted to assess how well the MLEs of the growth parameters fit the observed SFS. Put another way, if a researcher were to fit a growth model to the SFS from an admixed population, how likely is it that the researcher would reject the simple growth model as an explanation for the observed SFS? We performed a simple chi-square GOF test for a particular demographic model where we compared the observed SFS in each of the 500 simulated datasets to the expected SFS at the MLE estimates. Figure 3.8 shows a quantile-quantile (qq) plot comparing the GOF *P*-values from Pop AA versus those for Pop A. When $t_{cur} = 2400$, there is a shift toward smaller *P*-values in Pop AA as compared to Pop A (Figure 3.8). This effect is less pronounced when $t_{cur} = 4000$ (Figure 3.9). We find that when $t_{cur} = 2400$, 5% of the simulated datasets in Pop A have a *P*-value < 0.014. Note, the fraction of datasets with $P<0.014$ is greater than 1.4% due to the fact that some SNPs are linked, thus reducing the effective number of SNPs. Thus, we use 0.014 as an approximate 5% rejection region for the GOF test. Using this calibration, we find that 8.8% and 5.8% of datasets for the AA population have a *P*-value <0.014, for $N_{AA} = N_A$ and $N_{AA} =$

$0.1N_A$, respectively. When $t_{cur} = 4000$, 5% of the simulated datasets from Pop A have a P-value <0.0117 compared to 5.6% and 4.8%, for $N_{AA} = N_A$ and $N_{AA} = 0.1N_A$, respectively. These results suggest that there is a slightly worse GOF for the admixed population (Pop AA) than for the non-admixed population, but we cannot exclude the possibility that some of this pattern may be due to differences in how accurately we optimized the likelihood function across different models. Nevertheless, for the datasets simulated here (containing roughly 17,000 SNPs), the vast majority (91-95%) of datasets from Pop AA will be unable to reject the pure growth model. While we expect datasets from admixed populations containing more SNPs to have more power to reject the simplified (and incorrect) growth model, it appears that better model diagnostics are needed.

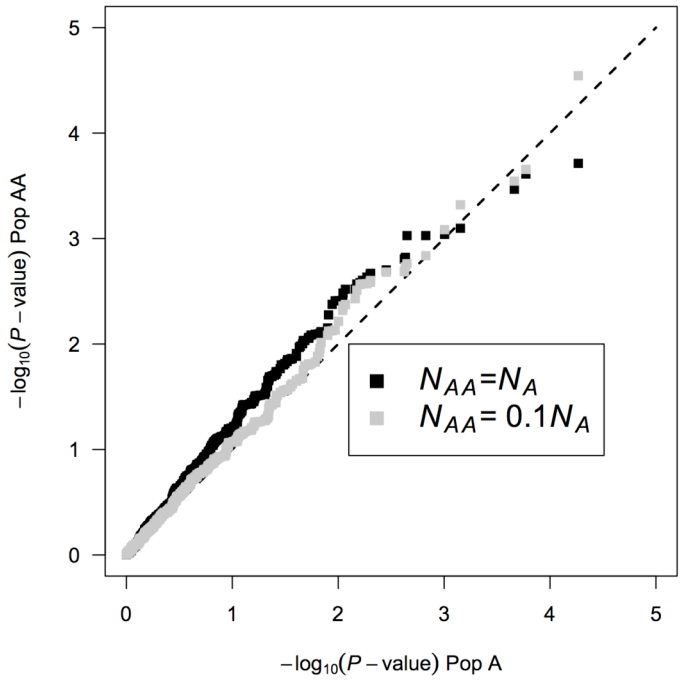


Figure 3.8: Quantile-Quantile (QQ) plot comparing the chi-square goodness of fit test *P*-values from data simulated from Pop A (*x*-axis) and Pop AA (*y*-axis). Note the excess of lower *P*-values in Pop AA relative to Pop A for both values of NAA. These results suggest that the best-fitting growth parameters tend to fit Pop AA (where the true demographic model involves admixture) slightly worse than they do for Pop A (where the true demographic model is a growth model). Here $t_{cur} = 2400$ generations.

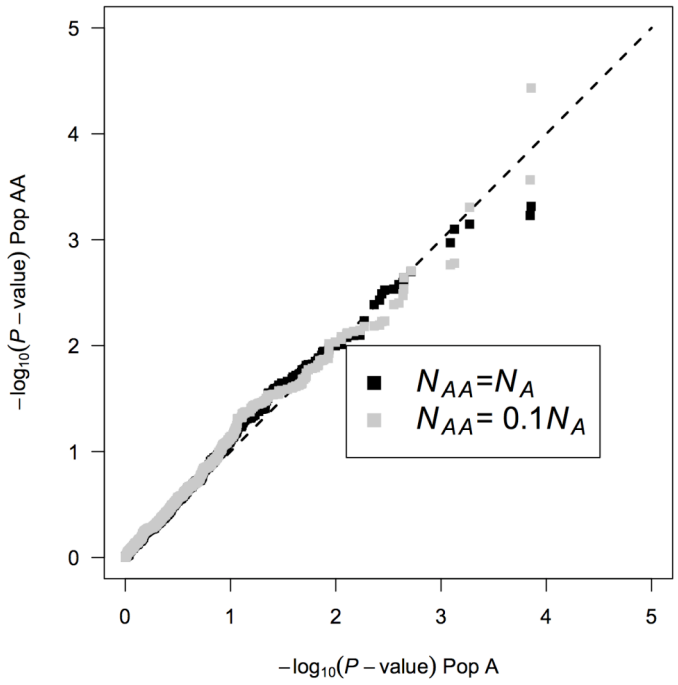


Figure 3.9: Quantile-Quantile (QQ) plot comparing the chi-square goodness of fit test *P*-values from data simulated from Pop A (*x*-axis) and Pop AA (*y*-axis). Note that there is not as much of an excess of low *P*-values for Pop AA as there was in Figure 3.8. Here t_{cur} = 4000 generations.

Inference of demography from human data

We estimated the three growth model parameters (t_{cur}, N_A, N_B/N_A) for the African American (AA) and Yoruba (YRI) populations using the SFS NIEHS resequencing dataset (Livingston *et al.* 2004). We chose to use the NIEHS dataset since it was generated by complete resequencing of the same genomic regions in both AA and YRI individuals. Since it is a complete resequencing dataset free of ascertainment bias, we can accurately estimate the SFS. Because the dataset included both AA and YRI individuals, we should be able to directly measure the effect that admixture has on estimates of the growth model parameters estimated from the SFS by comparing the parameter estimates of the two datasets. Importantly, since the same regions were studied and the resequencing was done by the same laboratory for the

two populations, any differences in the estimates should not be attributable to differences in selective pressure or laboratory errors.

As expected based on the analysis of simulated data described above, the folded SFS of the AA and YRI are fairly similar to each other ($P = 0.09$; $\chi^2 = 15.0155$; 9 df; Pearson's chi-square test), but the AA SFS has slightly more low frequency SNPs and more SNPs overall (see also Figure 3.10).

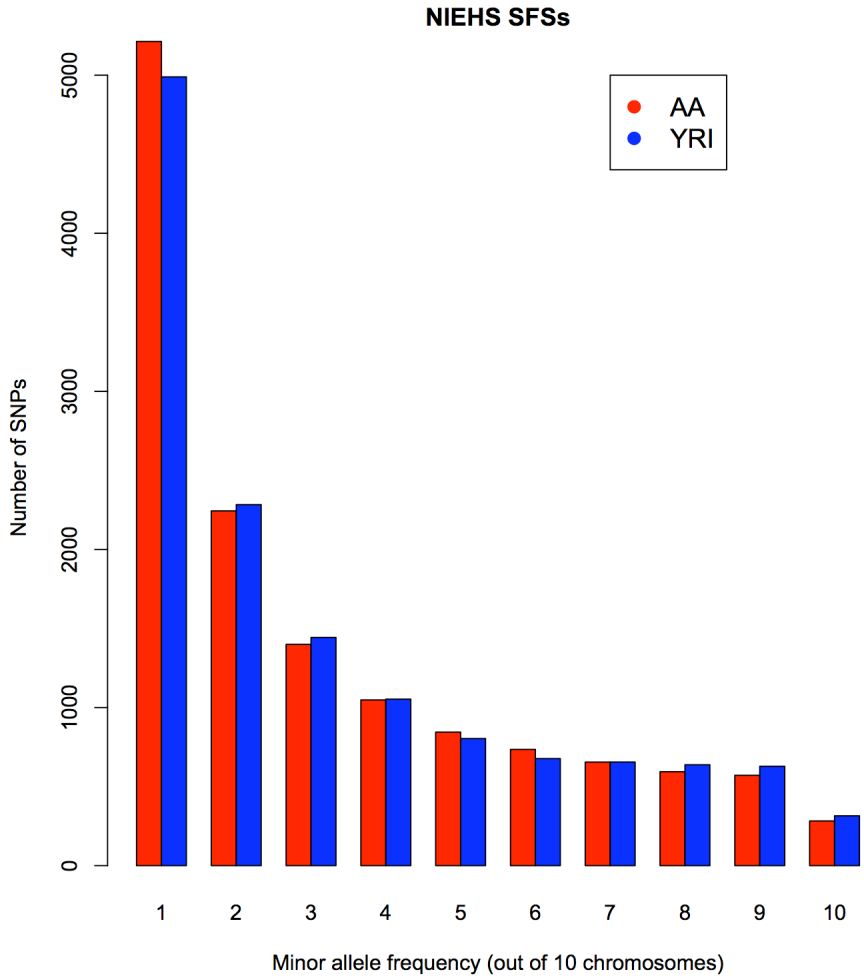


Figure 3.10: The folded SFS for the YRI and AA samples in the NIEHS dataset. The folded SFS presents the number of SNPs where the minor allele has a given frequency. Note, to allow for missing data, we projected the SFS to a sample size of 20 chromosomes.

Using the SNPs in the folded SFS, Wattersons's $\theta = 8.88 \times 10^{-4}$ per nucleotide in YRI and 8.95×10^{-4} per nucleotide in AA. The average number of pairwise differences (π)

per nucleotide is 7.99 x 10^{-4} in the YRI sample and 7.93 x 10^{-4} in the AA sample. We then estimated the three demographic parameters for the AA and YRI datasets (see Methods). Figure 3.11 shows the profile-likelihood curves for the three parameters. We find that the estimate of N_A is slightly higher in the AA population (15,732) as compared to the YRI population (14,647). N_B/N_A is slightly lower in the AA (0.46) than in the YRI (0.5) sample. The profile likelihood curves overlap substantially for t_{cur}, with a MLE of 5208 generations in the AA and 5425 generations in the YRI. Importantly, for all three parameters, the approximate 95% CIs from the profile-likelihood curves (<1.92 log-likelihood units) overlap between the AA and YRI estimates, suggesting that the YRI and AA parameter estimates are not significantly different from each other.

We then estimated the growth parameters for the Perlegen AA dataset (Hinds *et al.* 2005) using the HCN approach (see Methods). Figure 3.11 also shows the profile-likelihood curves for the three parameters. The estimate of N_A (12,500) is slightly smaller than the estimates obtained from the SFS based analyses. However, the estimates of the other two parameters using the HCN method are quite discordant with the estimates found using the SFS. N_B/N_A is much larger (0.95) when estimated using the HCN than the SFS (0.46). The timing of growth, t_{cur} is also estimated to be much more recent when using the HCN as compared to the SFS. Thus, as predicted by the analysis of simulated datasets described above, the HCN inference method gives different growth parameter estimates than the SFS inference method when the population truly had an admixed demographic history.

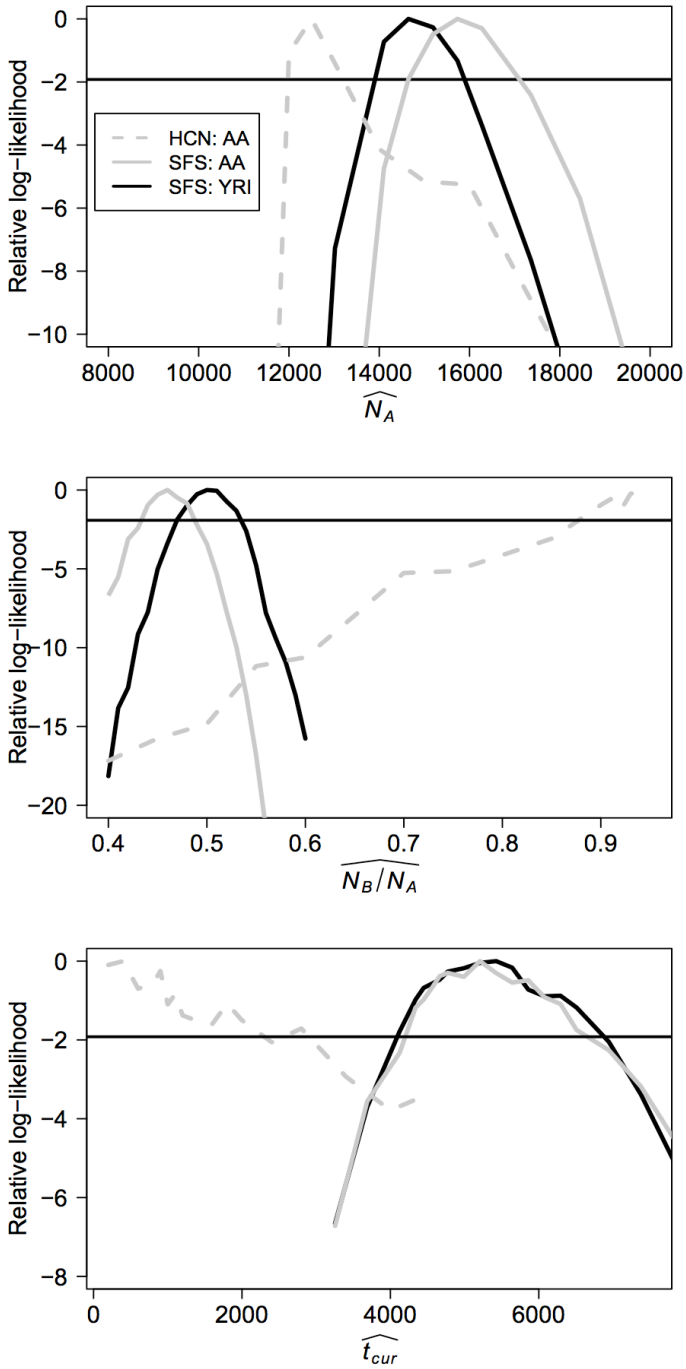


Figure 3.11: Profile log-likelihood curves for the three growth model parameters estimated from the NIEHS resequencing dataset (using the SFS for the AA and YRI samples) and the Perlegen SNP genotype dataset (using the HCN for the AA sample). Note that the two estimates based on the SFS (solid curves) are very similar to each other. The horizontal line in each figure denotes the approximate asymptotic 95% CI.

3.5 Discussion

We have examined how recent admixture affects estimates of population growth when using the SFS or the HCN statistic for inference. For certain parameter combinations, we find that growth parameter estimates made using the SFS in the admixed population are qualitatively similar to the true growth parameters from the un-admixed population. This pattern holds more often for the current population size (N_A) and growth parameter (N_B / N_A) than for the timing since growth parameter (t_{cur}), and seems to be little affected by whether or not the admixed population had experienced a reduction in size during its founding. If growth occurs at the time the ancestral populations split from each other (e.g. $t_{cur} = t_{split}$ in Figure 3.1), then estimates of t_{cur} from the admixed population will exhibit smaller bias.

We found that demographic inference based on haplotype patterns is more sensitive to the admixture process than demographic inference based on the SFS. Our simulations suggest that this difference comes about from the manner in which the 20% of ancestry from Pop E affects the SFS and HCN. The SFS from the admixed population (Pop AA) contains more SNPs and more low frequency SNPs than does the non-admixed population (Pop A). This is due to the fact that Pop E contains some population-specific SNPs not present in Pop A that are then brought into Pop AA during the admixture process. Based on the model assumed here, as well as the analysis of the NIEHS resequencing data, the extra SNPs brought into the admixed population from the European population do not substantially alter estimates of the population growth parameters. Conversely, the HCN statistic from the admixed population (Pop AA) is shifted towards a higher proportion of windows of the genome with fewer haplotypes and where the most common haplotype is at higher frequency than in the non-admixed Pop A. Essentially, this suggests that there is less haplotype diversity in Pop AA than in Pop A. This pattern arises because ~20% of

chromosomes in Pop AA are from Pop E, rather than from Pop A. Since Pop E has undergone a population bottleneck, it has less haplotype diversity than Pop A does. Consequently, Pop AA has lower haplotype diversity than Pop A simply because it contains ~20% of its chromosomes from the populations with lower haplotype diversity (Pop E) while Pop A contains 0% of its chromosomes from the population with lower diversity. Put another way, a single chromosome sampled from Pop A is more likely to represent a new haplotype in a sample from Pop A than a single chromosome sampled from Pop E would. This is the opposite of what was seen for single SNPs, where sampling chromosomes from a mixture of Pop A and Pop E results in an increase in the number of SNPs over sampling only Pop A (Ptak and Przeworski 2002; Stadler *et al.* 2009). This difference indicates that these two summaries of SNP data capture different and complementary aspects of ancestral history.

These finding offers some guidance for researchers wishing to infer demographic parameters in admixed populations. Due to the sensitivity of haplotype-based approaches to the admixture process, haplotype-based methods may be more informative than SFS-based methods for inferring the extent of recent admixture and detecting founder effects associated with admixture. However, when using haplotype-based approaches, researchers need to explicitly model the admixture process, rather than fitting a simplified growth model. As shown in the analysis of simulated and real data, fitting a simple growth model to an admixed population using haplotype patterns will likely give uninformative and erroneous results. Conversely, demographic inference using the SFS on admixed individuals often will yield parameter estimates for the non-admixed population that are likely to be qualitatively similar to those that would be obtained in the non-admixed population (at least for the range of parameters we have investigated here).

While existing computational methods use haplotype data for inference, these methods address different questions than those proposed here (Falush *et al.* 2003; Tang *et al.* 2006; Price *et al.* 2009). For example, the program STRUCTURE can be used to assign individuals to population clusters, estimate admixture proportions, and assign ancestry blocks (Falush *et al.* 2003). However, it assumes a simplified demographic model and does not allow inference of other demographic parameters, such as the timing and magnitude of population growth, or bottleneck parameters. Furthermore, STRUCTURE has mostly been used to detect admixture in recently admixed populations (less than 100 generations ago), rather than more ancient admixture events (Falush *et al.* 2003). A new method based upon the Li and Stephens copying model (Li and Stephens 2003) likely can detect older admixture events (Price *et al.* 2009). However, as currently implemented, like STRUCTURE, this approach does not provide estimates of population size changes. Thus, further work is needed to infer population size changes in admixed populations using haplotype patterns.

As predicted by our simulations, we found that the demographic parameters inferred using the SFS on the NIEHS AA are similar to those inferred from the NIEHS YRI individuals. This finding has implications for reconciling differences in estimates of population growth rates and times made using African and African American populations. Our finding from the NIEHS data of similar growth parameters for both the AA and YRI individuals suggests that using African American individuals as opposed to West African individuals should not lead to large differences in parameter estimates. Instead, we propose that the differences in estimates of growth parameters in different studies are likely to be due to differences in the amounts of natural selection in different datasets, systematic differences in laboratory protocols leading to different proportions in the SFS, or differences in modeling methods (e.g. whether or not migration is included). Consistent with this hypothesis, Wall *et al.* (2008) found

that a summary of the SFS, Tajima's *D* (Tajima 1989b), significantly differed among different datasets consisting of West African populations. Further studies using larger datasets with similar laboratory protocols and more advanced demographic models will help obtain more reliable parameter estimates.

While we examined complex demographic models involving population splits, bottlenecks, growth, and admixture, our models are still likely to be an over-simplification of the true demography of African and African American populations. Due to the many parameters in our model, we only examined a few illustrative examples, and did not evaluate systematically the effect of changing different parameters. For example, we assumed that the admixture event occurred 20 generations ago and that all Pop AA individuals have, on average, 80% of their ancestry from Pop A. In reality, both these parameters vary among individuals (Pfaff *et al.* 2001; Patterson *et al.* 2004). Our simulated datasets also assume a uniform recombination rate and do not include hotspots which are known to occur in the human genome (McVean *et al.* 2004). Thus, our models should be taken as illustrative examples of the effect of admixture on tests of demography and selection, rather than the full story. It is unclear if the general trends seen from our simulations, such as the finding that inference of demography using individuals from the recently admixed Pop AA provides qualitatively similar estimates obtained from individuals from the un-admixed parental population, Pop A, will hold under more complex models of demography. However, a more recent admixture time (7 generations instead of 20 generations) gave SFS that were qualitatively similar to those found when admixture occurred 20 generations ago (compare Figure 3.2 to Figure 3.3), suggesting that our conclusions may apply even if some of the true parameters differ slightly from those used in our models.

In our simulated data, Pop AA was made up of a mix of Pop A and Pop E where both these parental populations were assumed to be randomly mating. Genetic data suggest that African populations are highly structured (Reed and Tishkoff 2006; Campbell and Tishkoff 2008; Tishkoff *et al.* 2009; Bryc *et al.* 2010) and that multiple non-Bantu Niger Kordofanian-speaking populations likely have contributed ancestry to African Americans (Bryc *et al.* 2010). However, these non-Bantu Niger Kordofanian-speaking populations show very low levels of population differentiation with each other (Bryc *et al.* 2010), suggesting that our assumption of a randomly parental population for Pop AA may not be unreasonable.

Nevertheless, due to these inherent complexities in trying to jointly model African, African American, and European population history, we also analyzed empirical data from African and African American populations. The analysis of the NIEHS and Perlegen data allow us to test whether the predictions made from simulated data generated under our simplified demographic models hold for actual data generated under the true demographic model of these populations. The fact that the growth parameters estimated using the SFS from the NIEHS data in the Yoruba and African American populations were similar to each other, as predicted by our simulations, suggests that our simple models provide a reasonable guide to reality. Furthermore, the finding that the estimates of growth parameters using the HCN statistic African American data significantly differ from those estimated from the SFS is again consistent with the observations from our simulations.

While our analyses were done with the demography of West African and African American individuals in mind, our results suggest some general predictions for the future study of additional recently admixed populations, such as Latino populations. We should expect that the admixture process will have more severely altered haplotype patterns than the SFS, suggesting that haplotype-based approaches

should be more informative for learning about their recent demographic history. However, while haplotype patterns will provide useful information, they are also more sensitive to mis-specification of the demographic model. Thus, haplotype methods to estimate demographic parameters, like the HCN statistic, should not be used on admixed populations without properly modeling the admixture. Further complex demographic models and inference methods involving admixture will need to be developed to properly estimate model parameters.

Finally, we note that current model diagnostics (such as the Goodness-of-Fit statistic comparing observed to expected SFS) commonly used to assess concordance between model and SNP data, may not have power to detect when data were drawn from a more complex model. We suggest that LD and haplotype patterns as summarized by the HCN statistic are especially sensitive to recent admixture and may provide better diagnostics for detecting ill-fitting models. Pairwise LD patterns have recently been utilized by Hernandez *et al.* (2007b) and Gutenkunst *et al.* (2009) on large scale Macaque and human resequence data sets to assess the fit of the demographic models estimated from the SFS. A similar approach may provide a means of harmoniously and robustly undertaking model fitting and assessment for demographic inference from admixed populations as well.

CHAPTER 4

DETECTING DIRECTIONAL SELECTION IN THE PRESENCE OF RECENT ADMIXTURE[4]

4.1 Abstract

We investigate the performance of neutrality tests to detect selection in recently admixed populations. Tajima's *D*, Fu and Li's *D*, and haplotype homozygosity have lower power in the admixed population relative to the non-admixed population. Fay and Wu's *H* test, however, shows the opposite pattern. For the demographic models investigated, after accounting for the excess of low frequency alleles, ignoring recent admixture when defining the rejection regions of the frequency spectrum-based tests does not result in an excess of false-positive results.

4.2 Text

The classic model of genetic hitchhiking predicts that a favorable mutation which has recently fixed in the genome will be surrounded by reduced heterozygostiy (Maynard Smith and Haigh 1974; Kaplan *et al.* 1989), an excess of low frequency alleles (Braverman *et al.* 1995) and an excess of high frequency derived alleles (Fay and Wu 2000). These signatures have been used to detect the footprints of recent selection (reviewed in Nielsen 2005). However, the standard hitchhiking model assumes that selection acts on a newly arisen mutation in an additive manner in a random mating population of constant size. Recent work has examined the effect that violations of these assumptions have on patterns of variation in selected regions and on tests of neutrality (Kim and Stephan 2003; Innan and Kim 2004; Hermisson and Pennings 2005; Przeworski *et al.* 2005; Santiago and Caballero 2005; Kim 2006;

[4] Lohmueller, K.E., C.D. Bustamante, and A.G. Clark, Submitted.

Pennings and Hermisson 2006a; Pennings and Hermisson 2006b; Teshima and Przeworski 2006; Teshima *et al.* 2006; Thornton and Jensen 2007). Since violations of the assumptions of the standard hitchhiking model change the expected pattern of polymorphism around a selected site, it is important to characterize the effect of additional violations of the hitchhiking model.

Here we examine how recent admixture affects patterns of polymorphism around a recently selected site as well as the false-positive rates and power of common tests of neutrality. These topics are relevant to the interpretation of studies of positive selection in human populations that have studied American individuals with African and European ancestry (Akey *et al.* 2004; Carlson *et al.* 2005; Stajich and Hahn 2005; Kelley *et al.* 2006; Wang *et al.* 2006; Tang *et al.* 2007; Williamson *et al.* 2007; Nielsen *et al.* 2009). Furthermore, with the advent of next-generation sequencing data, it is anticipated that selection scans will be performed in additional recently admixed populations.

We simulated datasets with positive selection under the demographic model shown in Figure 4.1. Since many coalescent simulation programs that model positive selection do not allow for complex demographic models (Spencer and Coop 2004), we used the forward-simulation program SFSCODE (Hernandez 2008) to simulate datasets where positively selected mutations arose in Pop A. Our simulations differ from the standard coalescent models of selection since we introduce the selected allele at particular time points, rather than condition on some present day frequency of the selected allele (Braverman *et al.* 1995; Innan and Kim 2004; Spencer and Coop 2004). Since our simulations may include partial sweeps and sweeps that ended at different times, our simulations are not quantitatively comparable to other coalescent simulations of selective sweeps, though qualitative patterns should be similar.

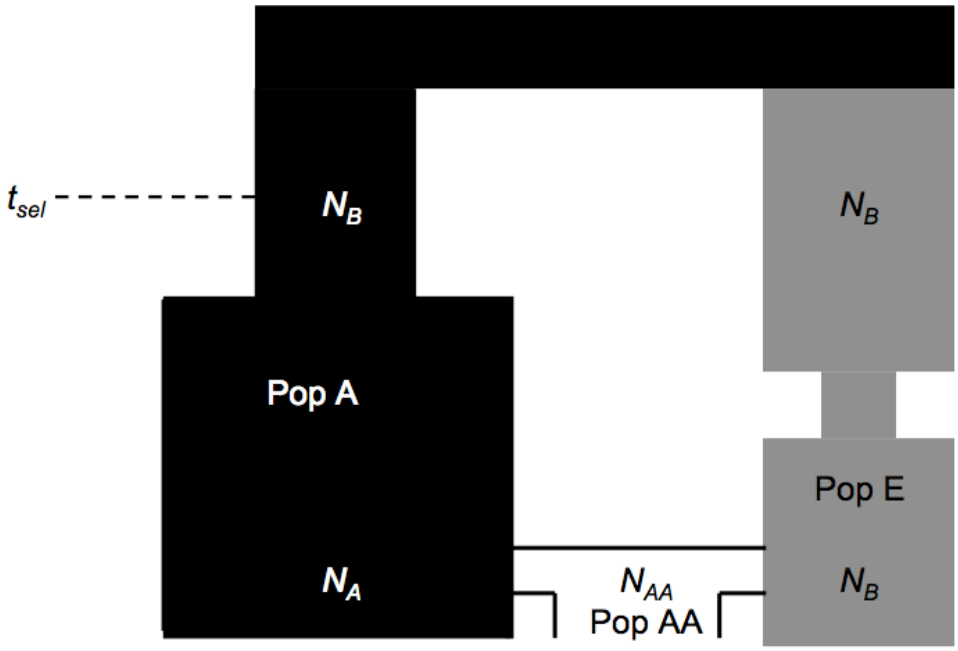
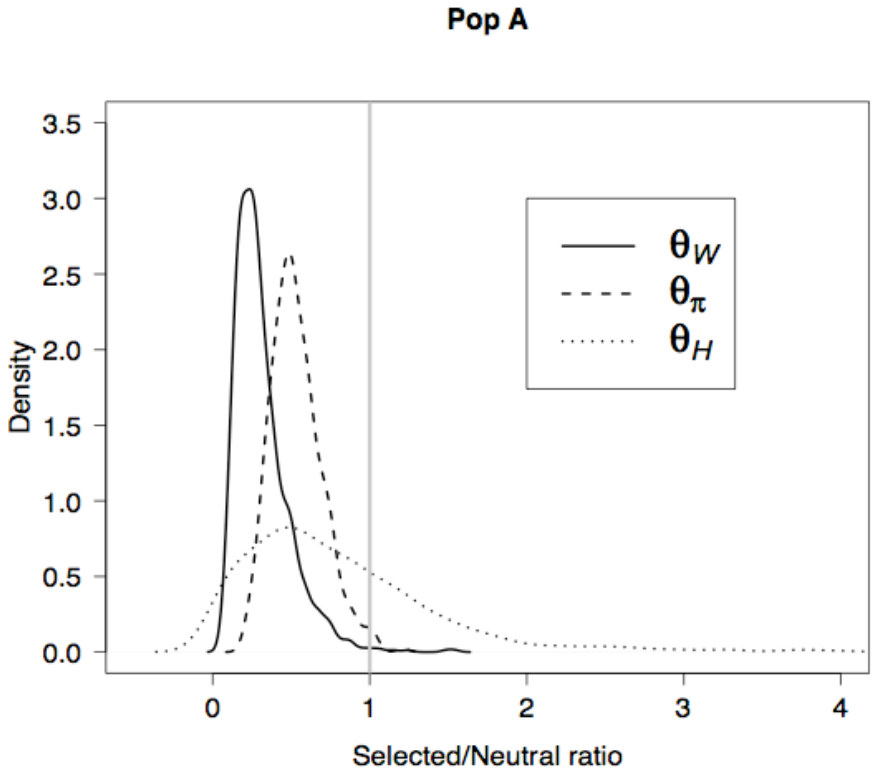


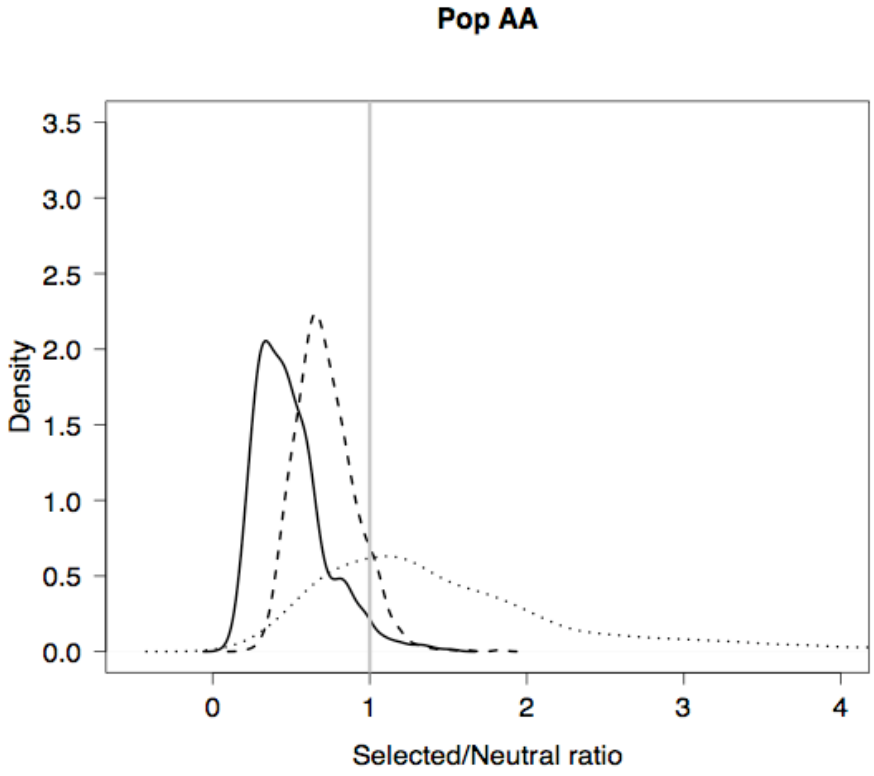
Figure 4.1: Demographic model used for simulations. This is the same model used in Lohmueller, Bustamante, Clark (*submitted*). 4000 generations ago the European (Pop E) and African (Pop A) populations split and (for simplicity) did not undergo any subsequent gene flow. The European population underwent a bottleneck (using parameters inferred by Lohmueller *et al.* 2009), with the exception of N_B, which is set to 10,000). The African American population (Pop AA) was formed 20 generations ago and has current size N_{AA} (Pfaff *et al.* 2001; Patterson *et al.* 2004; Tian *et al.* 2006). We assume 80% of the ancestry of Pop AA comes from Pop A, and 20% from Pop E (Pfaff *et al.* 2001; Patterson *et al.* 2004; Tian *et al.* 2006). Note, here N_A = 20,000; N_B/N_A = 0.5; t_{cur} = 2400 generations, and $N_{AA} = 0.1N_A$ or $N_{AA} = N_A$. Selected mutations occur in Pop A (black) for five generations, at time t_{sel}. All simulations assume an infinite sites mutation model and a Wright-Fisher model of reproduction.

Patterns of polymorphism around selected sites

Figure 4.2A shows the distribution of the ratio of diversity for regions that have undergone recent selection to neutrally evolving regions in the non-admixed population (Pop A).

Figure 4.2: Effect of recent admixture on patterns of variability around a selected site. The ratio of three different measures of diversity in regions containing a positively selected site to neutrally evolving regions. (A) The non-admixed population, Pop A, and (B) the admixed population, Pop AA. θ_π is the average number of pairwise differences between sequences (Tajima 1983), θ_W is Watterson's estimator based on the number of segregating sites (Watterson 1975), and θ_H is Fay and Wu's estimator based on high-frequency-derived alleles (Fay and Wu 2000). The proportion of the distribution to the left of the vertical line indicates a shift toward lower estimates of the measures of diversity in the selected regions relative to neutral regions. We simulated sequences containing a central 2 kb region where positively selected mutations could occur flanked by 25 kb of neutral sequence on either side in a sample size of 40 chromosomes. The population scaled selection coefficient, $\gamma = 2Ns = 100$, corresponding to moderate positive selection. We assumed an additive fitness model where an individual homozygous for the selected mutation has a fitness of 1+2*s* and heterozygotes have fitness 1+*s*. Since we only included positive selection in Pop A and not the other two populations, the central 2 kb region evolved neutrally for the entire simulation in Pop AA and Pop E. We set $\mu = r = 1 \times 10^{-8}$ per nucleotide which gives population scaled mutation ($\theta = 4N\mu$) and recombination ($\rho = 4Nr$) rates equal to 0.004 per-nucleotide. Since forward simulations are time consuming, we rescaled all population sizes to be two times smaller than those used in Figure 4.1, while keeping θ, ρ, and γ equal to their original values listed above. A similar strategy has been used in other forward simulations (Hoggart *et al.* 2007; Coop *et al.* 2009; Pickrell *et al.* 2009). Because SFSCODE cannot force a mutation to occur at a particular time, for five generations ending at time t_{sel} (1595-1600 generations ago, scaled by the smaller SFSCODE population size for Figures 4.2 and 4.3), all mutations that occurred in the central 2 kb regions were assigned a selection coefficient of $\gamma = 100$ and evolved under positive selection. Thus, the number of positively selected mutations per replicate is Poisson distributed with rate parameter $\lambda = 2$. All subsequent mutations that occurred in the central 2 kb regions were neutral. We retained only those replicates where the selected allele was not lost from the population. The statistics presented here are based on 1000 simulation replicates.


Pop A
Density
Selected/Neutral ratio
θW
θπ
θH


Pop AA
Density
Selected/Neutral ratio

Here we see the expected signatures of recent positive selection in the presence of recombination (Fu and Li 1993; Fay and Wu 2000; Nielsen 2005). All three estimators of diversity are reduced in selected regions relative to neutral regions (i.e. the bulk of the distributions are <1). Furthermore, θ_H is shifted toward higher values, indicating an excess of high-frequency-derived alleles, in some selected regions relative to neutral regions.

Figure 4.2B shows the same distributions for the admixed population (Pop AA). Again, there is an overall reduction in diversity in selected regions relative to neutral regions. However, both θ_π and θ_W are less reduced in selected regions in Pop AA than they were in Pop A. In fact, in Pop AA, some selected regions have higher values of θ_π and θ_W than neutrally evolving regions do. This pattern is due to variation from Pop E being brought into Pop AA during the admixture process. Also note that the ratio of θ_H in selected regions to neutrally evolving regions is higher in Pop AA than in Pop A, suggesting a more pronounced excess of high-frequency-derived variants around selected regions in Pop AA than in Pop A. An excess of high-frequency-derived alleles is expected to exist around a selected site while the selected allele is on its way to fixation, and shortly after the selected allele fixes in the population (Fay and Wu 2000; Przeworski 2002). However, many of the high-frequency-derived alleles present in the selected regions will become fixed in the population due to drift soon after the selected site fixes in the population. This is likely what happened in Pop A, since introducing the selected allele more recently, (~2400 generations ago, instead of ~3200 generations ago, assuming t_{sel} is scaled such that N_B = 10000) results in an increase of θ_H in Pop A (data not shown). Since we modeled population-specific selection in Pop A and not in Pop E, the ancestral alleles for many variants around the selected allele are still at high frequency in Pop E. When Pop AA is formed through mixing Pop A and Pop E, sites in the selected region where

the derived allele has fixed in Pop A now become polymorphic again and have the derived allele at high frequency, increasing θ_H in the admixed population. This mechanism predicts that θ_H in Pop AA should be less affected by the timing of the sweep than θ_H in Pop A would be. Indeed, our simulations show evidence of this. When the selected allele was introduced 2400 generations ago, rather than 3200 generations ago (scaled so that $N_B = 10000$), we find a higher increase of average θ_H in Pop A (5.13), than in Pop AA (0.93). Importantly, in both cases, average θ_H in Pop AA is larger than average θ_H in Pop A.

Performance of neutrality tests

We next examined the false positive and false negative rates for several common tests of neutrality in admixed vs. non-admixed populations. The purpose of this analysis was to address whether tests of neutrality have an elevated false-positive rate in admixed populations when admixture is not included in the null model to define the rejection region of test. Conversely, we wanted to assess whether admixed populations have higher or lower power at detecting population-specific positive selection in one of the parental populations than does using individuals from the parental population where the selective event occurred. For this analysis, we examined two different classes of test of neutrality: tests based on the site frequency spectrum (SFS; Tajima's *D* (Tajima 1989b), Fu and Li's *D* (Fu and Li 1993a), and Fay and Wu's *H* (Fay and Wu 2000)), and tests based on haplotype patterns (haplotype homozygosity).

Here we examine three different strategies to define rejection regions for each test. The quantiles of each statistic from neutral coalescent simulations were used to define the rejection region for each test. First, we used the standard neutral model (herein "SNM"). The SNM is anti-conservative for many of these tests when the true demographic model is more complex (Tajima 1989a; Tajima 1989b; Fu and Li 1993;

Simonsen *et al.* 1995; Przeworski 2002; Nielsen *et al.* 2005), and so the SNM serves as a baseline for comparison between admixed and non-admixed populations. The second strategy was to use the true demographic model (herein "TRUE") under which each population evolved. For Pop A, we used simulations from a growth model with the true growth parameters, and for Pop AA, we used simulations including admixture between Pop A and Pop E, again with the true model parameters. While the true demographic model is not known in practice, this strategy represents the best one could do with perfect demographic information. Finally, the third strategy was to estimate parameters of a simplified demographic model using the site frequency spectrum (SFS) from neutral data, and then do simulations under those parameters to define the rejection region (herein "EST"). This strategy mimics what is often done when researchers have genome-wide genetic variation data (Nielsen *et al.* 2009). Here we used a growth model for Pop A (where it is the correct model) as well as for Pop AA (where it is the incorrect model because the true model includes admixture).

Figure 4.3A shows the fraction of test statistics calculated from simulated neutral datasets that rejected neutrality for Pop A and Pop AA. As expected, we find an elevated false-positive rate for Tajima's *D* and Fu and Li's *D* in Pop A and Pop AA when using the rejection region defined by the SNM. This is caused by the excess of low frequency alleles present due to population growth (Tajima 1989a; Slatkin and Hudson 1991). When using the TRUE or EST rejection regions, this elevated false-positive rate disappears, because the excess of low-frequency alleles is not unusual under growth models. However, there is still a slight excess of false positive results in Pop AA when using the EST rejection region (~1%), although this pattern disappears when there is no founder effect in Pop AA ($N_{AA} = N_A$, Figure 4.4A). Fay and Wu's *H* test does not have an elevated false positive rate in Pop AA for any of the three rejection regions. This is reassuring, since in principle, if derived alleles fix in Pop A,

but not in Pop E, then the admixed population (Pop AA) could contain derived alleles at high frequency which would yield false-positive results (Fay and Wu 2000). However, this scenario appears to be uncommon for neutral data under the demographic model used here simply due to the fact that the level of differentiation between Pop A and Pop AA is not high enough to result in many cases where the derived allele is at near-fixation in one population but low frequency in the other. A simulation study (Przeworski 2002) that suggested population structure coupled with little migration can be a potential source of false-positives for this test assumed a two-island model with higher levels of population differentiation than those considered here. Our simulations also suggest that the strategy used by Nielsen *et al.* (2009), where growth model parameters are estimated using the neutral SFS and then simulations with those parameter estimates are used to define the rejection region for neutrality tests based on the SFS, is a reasonable strategy that does not result in an excess of false-positives, even when the true demographic model involves both population growth and admixture. Of course, this result only holds if neutral regions of the genome can identified and used for demographic inference.

The haplotype homozygosity test appears to be conservative in both Pop A and Pop AA when using the SNM to define the rejection region, presumably due to the fact that the SNM assumes a smaller current population size than do the growth models, and thus a smaller population-scaled recombination rate. When using TRUE rejection regions, the test behaves appropriately. However, we note a slight excess of false-positive results (~3% excess) when using the EST demographic model for Pop AA. This is likely due to an excess of haplotype homozygosity from the admixture or founding event that is unexpected under a simple growth model fit using the SFS. We did not find an elevated false-positive rate when $N_{AA} = N_A$ (Figure 4.4), suggesting that

this pattern is more due to the population size reduction at the time of admixture, rather than the admixture process itself.

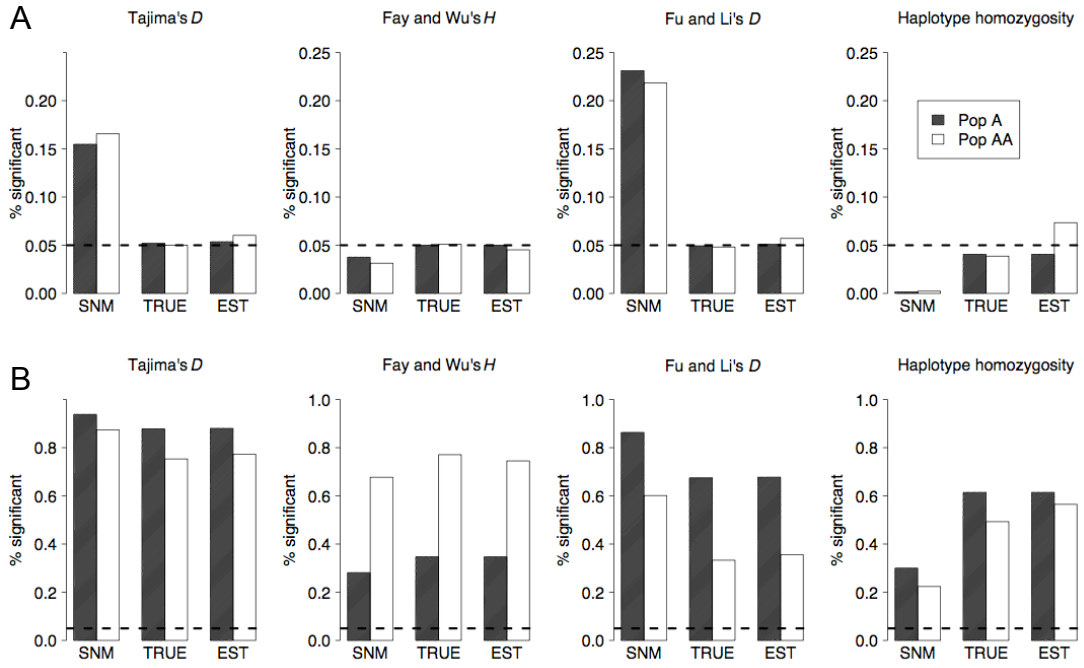


Figure 4.3: Performance of neutrality tests in admixed populations. (A) Proportion of neutral datasets rejecting neutrality (False-positives) for different rejections regions. (B) Proportion of selected datasets rejecting neutrality (Power) for different rejection regions. SNM denotes the rejection region defined by the standard neutral model, TRUE the rejection region defined by the true demographic model for each population, and EST the rejection region defined by a growth model where the parameters were estimated from the SFS of neutral data (see text). The parameters of the growth model were those at the modes of the distributions of MLEs estimated from the SFS for the growth parameters in Figure 3.4 of Lohmueller, Bustamante, Clark (*submitted*). The same datasets simulated with selection in Figure 4.2 were used here.

Thus, researchers should be cautious when using a demographic model inferred from the SFS as a null model for haplotype-based tests of neutrality. The problem might be circumvented by either using more accurate demographic models or inferring demographic models using haplotype-based approaches (Lohmueller *et al.* 2009).

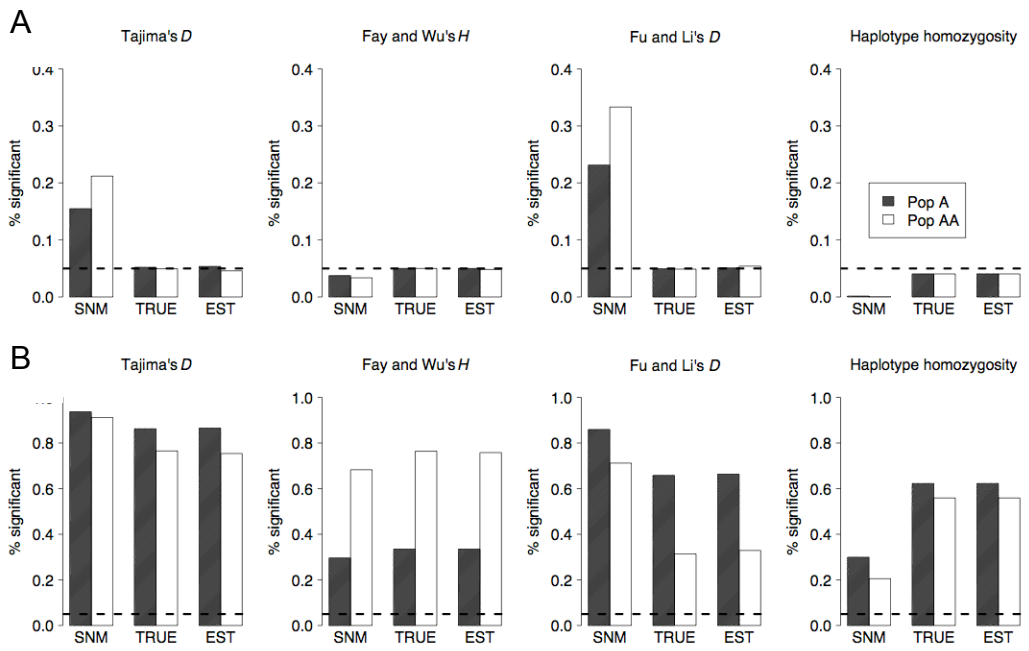


Figure 4.4: Performance of neutrality tests in admixed populations when $N_{AA} = N_A$. (A) Proportion of neutral datasets rejecting neutrality (False-positives) for different rejections regions. (B) Proportion of selected datasets rejecting neutrality (Power) for different rejection regions. SNM denotes the rejection region defined by the standard neutral model, TRUE the rejection region defined by the true demographic model for each population, and EST the rejection region defined by a growth model where the parameters were estimated from the SFS of neutral data (see text). All demographic parameters are the same as in Figure 4.3.

We next assessed the power of these neutrality tests to detect positive selection that occurred only in Pop A when using individuals sampled from Pop A and Pop AA. Figure 4.3B shows the fraction of tests using datasets simulated with positive selection that rejected neutrality for Pop A and Pop AA. Here we find that for Tajima's *D*, Fu and Li's *D*, and haplotype homozygosity, for all three rejection regions, individuals from Pop AA have lower power to detect selection than individuals from Pop A. For Tajima's *D*, this pattern can be explained by the admixture process increasing θ_π more than it increases θ_W (Figure 4.2B). Increasing θ_π more than θ_W will lead to larger values of *D*, and consequently, lower power to detect selection. We find that Fay and Wu's *H* test has nearly twice as much power to detect selection using the admixed population (Pop AA) than the non-admixed population (Pop A) for all three rejection

regions. This increase in power is due to the large increase in θ_H relative to θ_π in selected regions in Pop AA. We should point out that this result is likely sensitive to the timing and strength of selection simulated and may not apply universally for all types of population-specific selection. However, we find that all of these results qualitatively hold when the selected allele is introduced at different times (Figure 4.5).

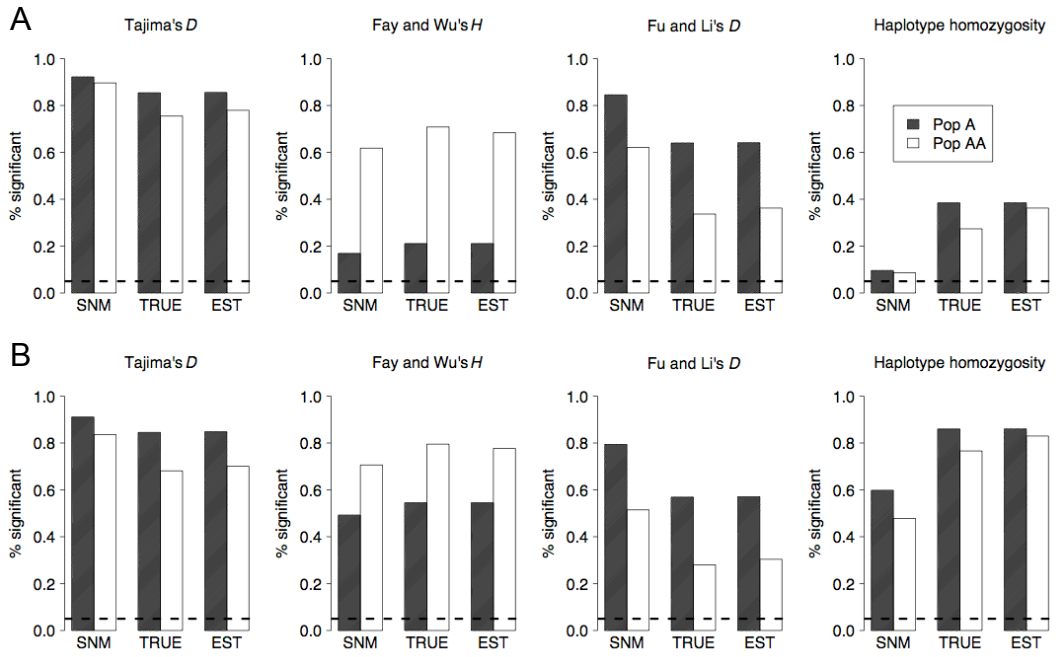


Figure 4.5: Proportion of selected datasets rejecting neutrality (Power) when the selected mutations occurred at different times. (A) The selected mutations occurred when Pop A and Pop E split (t_{sel} = 4000 generations when N_B is scaled to equal 10000). (B) The selected mutations occurred when Pop AA expands (t_{sel} = 2400 generations when N_B is scaled to equal 10000). SNM denotes the rejection region defined by the standard neutral model, TRUE the rejection region defined by the true demographic model for each population, and EST the rejection region defined by a growth model where the parameters were estimated from the SFS of neutral data (see text). All demographic parameters are the same as in Figure 4.3.

Conclusions

Our simulations are meant to be illustrative examples of the effect of admixture on patterns of variation in selected regions. Specific quantitative predictions are likely to be dependent on the parameters of the selection model used, the timing of the admixture event(s), and the overall demographic model employed.

Nevertheless, these simulations can help interpret previous studies of selection in admixed populations and suggest strategies for detecting selection in admixed populations.

Studies that have found evidence of positive selection using frequency spectrum-based tests of neutrality in African American populations (Akey *et al.* 2004; Carlson *et al.* 2005; Stajich and Hahn 2005; Kelley *et al.* 2006; Williamson *et al.* 2007) are not likely to have an elevated false-positive rate due to recent European admixture. Our simulations partially support the hypotheses of previous studies (Carlson *et al.* 2005; Voight *et al.* 2006; Williamson *et al.* 2007) that recent European admixture in African American populations can obscure some of the signal of recent selective sweeps in African populations. However, recent admixture accentuates signatures of selection based on high-frequency derived variants, suggesting there still can be reasonable power to detect recent selection in admixed populations.

CHAPTER 5

PROPORTIONALLY MORE DELETERIOUS GENETIC VARIATION IN EUROPEAN THAN AFRICAN POPULATIONS[5]

5.1 Abstract

Quantifying the number of deleterious mutations per diploid human genome is of critical concern to both evolutionary and medical geneticists (Muller 1950; Fay *et al.* 2001; Cohen *et al.* 2004). Here, we combine genome-wide polymorphism data from PCR-based exon re-sequencing, comparative genomic data across mammalian species, and protein structure predictions to estimate the number of functionally consequential mutations carried by each of 15 African American (AA) and 20 European American (EA) individuals. We find that AAs show significantly higher levels of nucleotide heterozygosity than do EAs for all categories of functional mutations considered including synonymous, nonsynonymous, predicted "benign", predicted "possibly damaging" and predicted "probably damaging" mutations. This result is wholly consistent with previous work showing higher overall levels of nucleotide variation in African populations as compared to Europeans (Tishkoff and Williams 2002). EA individuals, on the other hand, have significantly more genotypes homozygous for the derived allele at synonymous and nonsynonymous SNPs and for the damaging allele at "probably damaging" SNPs than AAs do. Surprisingly, for SNPs segregating only in one population or the other, the proportion of nonsynonymous SNPs is significantly higher in the EA sample (55.4%) than in the AA sample (47.0%; $P < 2.3 \times 10^{-37}$). We observe a similar proportional excess of SNPs that are inferred to be "probably damaging" (15.9% EA; 12.1% AA; $P < 3.3 \times$

[5] Previously published in Lohmueller *et al*. (2008) and has been reproduced with permission. Copyright the Authors.

10^{-11}). Using extensive simulations, we show that this excess proportion of segregating damaging alleles in Europeans is likely a consequence of a bottleneck that Europeans experienced around the time of the migration out of Africa.

5.2 Main text

Current estimates of the number of deleterious mutations per diploid human genome vary by several orders of magnitude. Using a correlation in inbreeding rates within consanguineous marriages and mortality, Morton, Crow, and Muller (Morton *et al.* 1956) estimated each of us carries 3-5 lethal equivalents (*i.e.*, an allele or combination of alleles that if made homozygous would be lethal) whereas Kondrashov (Kondrashov 1995) has predicted that the number may be as high as 100 lethal equivalents. Comparative genomic methods suggest that approximately 38% of amino-acid changing polymorphisms are deleterious, with 1.6 new deleterious mutations arising per individual per generation (Eyre-Walker and Keightley 1999) while studies based on segregating polymorphisms estimate that each person carries between 500 and 1,200 deleterious mutations (Fay *et al.* 2001; Sunyaev *et al.* 2001). It is very difficult to reconcile these estimates since each study used different methods and data. Furthermore, studies that used DNA sequences only included data from several hundred genes. Thus, there is a critical need for an unbiased genome-wide estimate of the number of damaging mutations carried by individuals in different populations.

We quantify the number of damaging mutations per diploid human genome by combining the Applera genome-wide survey of SNPs found by resequencing of 20 European Americans (EAs) and 15 African Americans (AAs; Bustamante *et al.* 2005) with comparative genomic data including the PanTro2 build of the chimpanzee genome and protein structure prediction data. After applying strict quality control

criteria, the data set we analyzed contains 39,440 autosomal SNPs free of ascertainment bias comprising 10,150 unique transcripts in the human genome (see Methods). Of these SNPs, 20,893 were synonymous (nucleotide changes that do not change the amino acid) and 18,547 were nonsynonymous (nucleotide changes that change the amino acid).

At each SNP, an individual can be homozygous for the ancestral allele (carry zero copies of the mutant allele), heterozygous (carry one copy of the mutant allele), or homozygous for the derived allele (carry two copies of the mutant allele). We find that an individual is heterozygous, on average, for 1,962.4 nonsynonymous SNPs (SD: 275.1; Figure 5.1a; Table 5.1). These numbers are an underestimate since only SNPs with good quality sequence and a matching chimp base are considered. Perhaps for these reasons, our estimate is slightly smaller than that by Cargill *et al.* (1999), even after decreasing their estimate to account for the current estimated number of genes in the genome. For both synonymous and nonsynonymous SNPs, AA individuals are heterozygous at a greater number of SNPs than are EA individuals (Figure 5.1a; $P < 6.2 \times 10^{-10}$, Mann-Whitney U-test (MWU) for synonymous SNPs; $P < 6.2 \times 10^{-10}$, MWU for nonsynonymous SNPs), consistent with previous studies finding higher levels of genetic variability in Africa (Tishkoff and Williams 2002). Interestingly, for both types of SNPs, we find that EA individuals are homozygous for the derived allele at a greater number of SNPs than AA individuals (Figure 5.1b; $P < 6.2 \times 10^{-10}$, MWU). These patterns are largely due to an elevated number of SNPs fixed for the derived allele in the EA sample while segregating for two alleles in the AA sample. Excluding SNPs that are not segregating in the particular sub-population, we observe that AAs have more homozygous derived genotypes per individual at synonymous SNPs and EAs slightly more homozygous derived genotypes per individual at nonsynonymous SNPs.

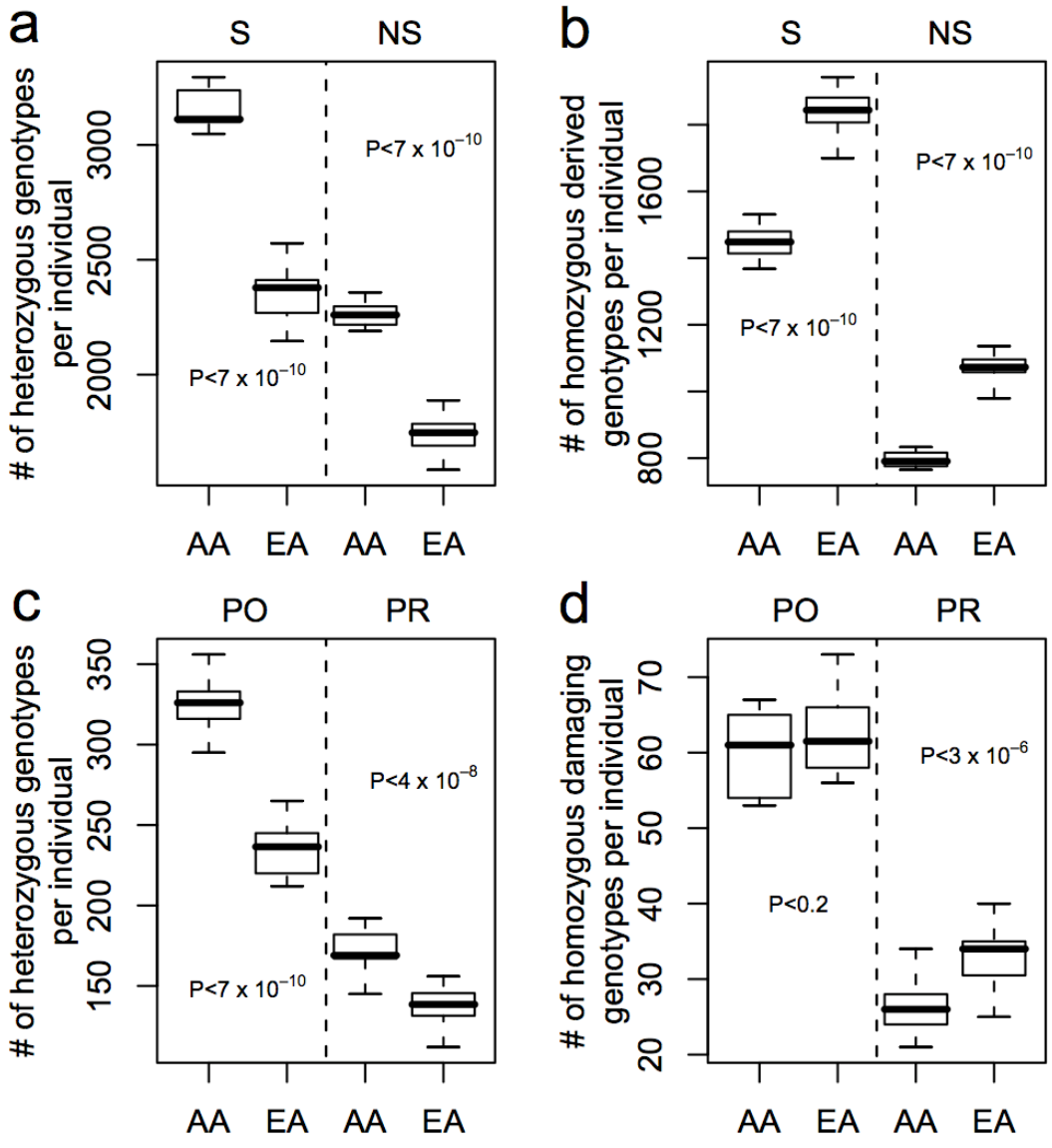


Figure 5.1: Distribution of the number of heterozygous and homozygous genotypes per individual a, Number of heterozygous genotypes per individual at synonymous (S) or nonsynonymous (NS) SNPs. **b,** Number of genotypes homozygous for the derived allele per individual at synonymous (S) or nonsynonymous (NS) SNPs. **c,** Number of heterozygous genotypes per individual at possibly damaging (PO) or probably damaging (PR) SNPs. **d,** Number of genotypes homozygous for the damaging allele at possibly damaging (PO) or probably damaging (PR) SNPs. Dark horizontal lines within boxes indicate medians, and the whiskers indicate the ranges of the distributions. EA: European American; AA: African American.

Table 5.1 Number of SNPs per individual.

Individual	# het S	# hom der S	Total # gtps S	# het NS	# hom der NS	total # gtps NS	# het B	# hom dam B	total #gtps B	# het PO	# hom dam PO	total # gtps PO	# het PR	#hom dam PR	Total # gtps PR
EA															
NA00131	2255	1822.3	18922	1586	1054.2	16804	861	769	8873	212	58	3034	141	31	1899
NA00333	2281	1942.5	19222	1656	1121	17151	913	808	9052	217	72	3088	132	36	1947
NA00546	2420	1807.5	19293	1784	1066.6	17124	997	779	9039	231	65	3078	135	34	1930
NA00607	2392	1864.6	19515	1741	1060.8	17316	945	781	9132	244	56	3124	152	29	1964
NA00893	2504	1834.1	19747	1873	1069.3	17564	1033	792	9249	238	64	3180	137	32	1989
NA00946	2404	1852.2	19497	1795	1063.6	17363	1001	757	9129	257	67	3150	154	34	1974
NA01805	2257	1923.8	19386	1754	1110.6	17251	948	798	9110	250	68	3109	146	33	1941
NA01814	2184	1897.9	19117	1627	1096.4	16943	886	810	8953	218	63	3067	122	39	1936
NA01953	2392	1903.2	19331	1787	1054.5	17226	1001	791	9088	231	58	3116	129	30	1951
NA01954	2378	1860.4	19408	1803	1090.5	17325	966	810	9142	243	71	3134	144	35	1956
NA01990	2396	1864.8	19517	1785	1066.7	17390	973	775	9189	238	61	3131	151	34	1966
NA02254	2516	1806.2	19539	1768	1110.1	17389	980	811	9185	218	61	3133	156	35	1949
NA05920	2351	1920.1	19583	1761	1076.3	17414	966	801	9175	230	62	3168	145	30	1973
NA08587	2572	1770.1	19431	1888	990.4	17278	1016	755	9083	265	58	3135	131	34	1975
NA09947	2215	1790.3	18522	1691	1033.9	16539	914	756	8725	237	58	2997	134	25	1862
NA10924	2146	1699.6	17703	1602	979.5	15787	877	698	8284	214	60	2880	131	31	1772
NA10959	2290	1824.9	19054	1691	1092.2	16958	903	796	8964	246	57	3058	134	34	1914
NA12593	2341	1790.7	19023	1707	1095.6	16870	948	799	8906	222	63	3050	140	37	1915
NA14448	2460	1842.1	19392	1714	1076.9	17314	926	812	9150	248	59	3128	112	30	1961
NA14548	2379	1845.5	19417	1697	1135.9	17296	940	807	9151	236	73	3125	142	40	1944
Mean	2356.7	1843.1	19231	1735.5	1072.3	17115.1	949.7	785.3	9029	234.8	62.7	3094.3	138.4	33.2	1935.9

Notes: het=Number of heterozygous genotypes for that individual; hom der=Number of genotypes where the individual was homozygous for the derived allele. We corrected for ancestral misidentification (see Methods) which is why decimal numbers are possible; total # gtps=Total number of genotypes for that individual; S=synonymous; NS=nonsynonymous; B=Benign; PO=Possibly damaging; PR=Probably damaging. For the PolyPhen classes, here we only count SNPs where the derived allele matches the predicted damaging allele. The damaging allele is the less-conserved one in the multi-alignment and has the lower PSIC score.

Table 5.1 (Continued)

Individual	# het S	# hom der S	Total # gtps S	# het NS	# hom der NS	total #gtps NS	#het B	#hom dam B	Total #gtps B	#het PO	#hom dam PO	Total # gtps PO	#het PR	#hom dam PR	Total # gtps PR
AA															
NA14439	3244	1369.1	19349	2358	777.2	17223	1276	618	9086	344	55	3133	190	24	1936
NA14454	3112	1425.9	19361	2250	822.7	17282	1258	645	9116	295	61	3126	169	34	1944
NA14464	3100	1426.6	19265	2260	768.2	17089	1273	622	9022	327	53	3083	169	24	1929
NA14476	3099	1391.4	18910	2190	767.6	16927	1186	616	8913	319	61	3071	167	28	1890
NA14480	3107	1401.9	19288	2243	817.3	17076	1175	668	9005	356	53	3068	167	25	1958
NA14501	3232	1448.6	19380	2209	774.6	17088	1233	629	9034	323	53	3059	152	29	1928
NA14503	3150	1468.6	19587	2276	815.3	17447	1289	644	9207	317	67	3143	182	22	1958
NA14508	3277	1442.1	19644	2297	790.6	17469	1261	633	9228	329	66	3159	182	28	1963
NA14511	3276	1530.8	19819	2226	833.3	17527	1214	656	9264	311	66	3149	167	24	1985
NA14535	3048	1448.3	19140	2200	784.7	16920	1195	629	8953	330	56	3047	145	26	1903
NA14632	3050	1468.3	19280	2353	790.6	17188	1265	648	9106	339	53	3070	192	30	1941
NA14649	3196	1496.4	19490	2342	815.7	17366	1264	665	9193	336	64	3129	188	27	1949
NA14663	3295	1494.9	19768	2283	825.1	17552	1266	665	9263	326	57	3172	171	25	1973
NA14665	3095	1491	19508	2298	805.7	17340	1243	652	9144	315	67	3127	178	27	1957
NA14672	3102	1367.6	18579	2190	765.1	16509	1227	627	8749	306	64	2972	158	21	1862
Mean	3158.9	1444.8	19357.9	2265	796.9	17200.2	1241.7	641.1	9085.5	324.9	59.7	3100.5	171.8	26.3	1938.4
MWU *P*-value	$<7x10^{-10}$	$<7x10^{-10}$	0.63	$<7x10^{-10}$	$<7x10^{-10}$	0.73	$<7x10^{-10}$	$<7x10^{-10}$	0.53	$<7x10^{-10}$	0.16	0.84	$<4x10^{-8}$	$<3x10^{-6}$	0.7

Notes: het=Number of heterozygous genotypes for that individual; hom der=Number of genotypes where the individual was homozygous for the derived allele; We corrected for ancestral mis-identification (see Methods) which is why decimal numbers are possible; total # gtps=Total number of genotypes for that individual; S=synonymous; NS=nonsynonymous; B=Benign; PO=Possibly damaging; PR=Probably damaging; For the PolyPhen classes, here we only count SNPs where the derived allele matches the predicted damaging allele. The damaging allele is the less-conserved one in the multi-alignment and has the lower PSIC score. MWU *P*-value is a 2-sided P-value from the Mann-Whitney U test comparing the column of interest in EA and AA.

To estimate the number of damaging alleles carried by each individual in our sample, we used the PolyPhen algorithm (Sunyaev *et al.* 2001; Ramensky *et al.* 2002) to predict which nonsynonymous SNPs might disrupt protein function. PolyPhen predicts whether a SNP is "benign", "possibly damaging", or "probably damaging" based on evolutionary conservation and structural data. In order to assess whether "damaging" SNPs were more likely to be deleterious, we compared the allele frequency distribution of SNPs predicted to be "benign", "possibly damaging", and "probably damaging" for each population. We find that the three distributions are significantly different from each other, with more low frequency SNPs in the "probably damaging" category (Table 5.2, $P < 5.9 \times 10^{-81}$ AA, $P < 2.3 \times 10^{-101}$ EA, Kruskal-Wallis test), suggesting that the majority of SNPs classified as damaging are also evolutionarily deleterious.

Table 5.2: Distribution of Applera SNPs by population and functional class.

Category	Shared	Private AA	Private EA	Mean derived frequency AA[1]	Mean derived frequency EA[2]
Synonymous (%)	8,056 (58.3%)	8,958 (53.0%)	3,879 (44.6%)	0.211	0.266
Nonsynonymous (%)	5,771 (41.7%)	7,950 (47.0%)	4,826 (55.4%)	0.174	0.202
Benign (%)	4,448 (78.6%)	5,260 (67.7%)	2,928 (62.1%)	0.200	0.238
Possibly damaging (%)	795 (14.0%)	1,572 (20.2%)	1,035 (22.0%)	0.113	0.119
Probably damaging (%)	422 (7.4%)	942 (12.1%)	749 (15.9%)	0.099	0.108

[1]Average frequency using SNPs segregating in the AA sample. No correction for ancestral mis-identification was used.
[2]Average frequency using SNPs segregating in the EA sample. No correction for ancestral mis-identification was used.

Figure 1c-d shows the distribution of the number of SNPs per individual where individuals were heterozygous (Figure. 5.1c) and homozygous for the damaging allele

(Figure 5.1d) for SNPs predicted to be "possibly damaging" and "probably damaging". We find that an individual typically carries 426.1 damaging (here defined as possibly or probably damaging) SNPs in the heterozygous state (SD: 65.4, range: 340-534) and 91.7 in the homozygous state (SD: 8.6, range: 77-113). Since we surveyed just over 10,000 genes, the actual number of damaging mutations in a person's genome may be as much as twice that given here. Every individual in our sample is heterozygous at fewer "probably damaging" SNPs than synonymous SNPs, consistent with purifying selection eliminating damaging SNPs from the population. AAs have significantly more heterozygous genotypes than do EAs for all three PolyPhen categories (Figure 5.1c, $P < 6.2 \times 10^{-10}$, for "possibly damaging" SNPs; $P < 3.7 \times 10^{-8}$, for "probably damaging" SNPs). The two populations differ significantly in the distribution of homozygous genotypes for the damaging allele at "probably damaging SNPs" (Figure 5.1d; $P < 2.7 \times 10^{-6}$), with EAs having approximately 26% more homozygous damaging genotypes than AAs. The lack of a statistical difference at "possibly damaging" SNPs ($P = 0.17$) is likely due to a lack of power since, overall, all other categories of SNPs (synonymous, nonsynonymous, "benign", and "probably damaging") follow the same pattern of excess homozygosity for the derived/damaging allele in EAs relative to AAs.

Classical analyses of human inbreeding suggest that each individual carries 1.44-5 lethal equivalents (Morton *et al.* 1956; Bittles and Neel 1994). However, inbreeding studies cannot determine whether a single lethal equivalent is due to one lethal allele, two alleles each with a 50% chance of lethality, 10 alleles each with a 10% chance of lethality, or other combinations. Since we find that individuals carry hundreds of damaging alleles, it is likely that each lethal equivalent consists of many weakly deleterious alleles. Our finding that each person carries several hundred

potentially damaging SNPs suggests that large-scale medical re-sequencing will be useful to find common and rare SNPs of medical consequence (Cohen *et al.* 2004).

We next examined the distribution of synonymous and nonsynonymous SNPs between AA and EA population samples (Table 5.2). As expected (Tishkoff and Williams 2002), there are more of both types of SNPs in the AA sample than in the EA sample. However, when classifying synonymous and nonsynonymous SNPs as being shared, private to AA, or private to EA, we strongly reject homogeneity (Table 5.3, $P < 3.0 \times 10^{-88}$).

Table 5.3: Results of G-tests of homogeneity for Table 5.2.

	Nonsynonymous vs. Synonymous			**Benign vs. Possibly vs. Probably damaging**		
	G	**df**	***P*-value**	***G***	**df**	***P*-value**
Shared vs. private AA vs. private EA	403.1	2	3.0×10^{-88}	377.8	4	1.8×10^{-80}
Shared vs. Private	239.9	1	4.3×10^{-54}	329.5	2	2.9×10^{-72}
Private AA vs. Private EA	163.2	1	2.3×10^{-37}	48.3	2	3.3×10^{-11}

We find the proportion of private SNPs that are nonsynonymous (49.9%) is higher than the proportion of shared SNPs that are nonsynonymous (41.7%; $P < 4.3 \times 10^{-54}$), which is not surprising since nonsynonymous SNPs are more likely to be at lower frequency and thus be population specific. However, considering only the private SNPs, we find that the EA sample has a higher proportion of nonsynonymous SNPs (55.4%) than the AA sample (47.0%; $P < 2.3 \times 10^{-37}$). We observed a similar significant proportional excess of private nonsynonymous SNPs in an independent data set collected by the SeattleSNPs project (Table 5.4; Supplementary Note 1 in APPENDIX 2). The SeattleSNPs data, additional quality control analyses (Supplementary Note 2 in APPENDIX 2 and Table 5.5), and a similar finding reported for the *ANGPTL4* locus (Romeo *et al.* 2007) indicate that this pattern is not an artefact of the Applera data. Our further analyses using Yoruba individuals from Nigeria

collected by the International HapMap Consortium (International HapMap Consortium 2007), support this result indicating that it is robust to admixture (Supplementary Note 3 in APPENDIX 2).

Table 5.4: Distribution of SNPs between populations: SeattleSNPs p1 and p2

Data set	Count Shared	Count Private AA	Count Private EA	Total Table[1]	Shared vs. Private[2]	AA private vs. EA private[3]
Seattle SNPs p1						
Synonymous	198	237	59			
Nonsynonymous	186	247	118			
% Nonsynonymous	48.4%	51.0%	66.7%			
Significance				G=17.5 $P<1.6\times10^{-4}$ 2 df	G=4.5 $P<0.035$ 1 df	G=13.0 $P<3.1\times10^{-4}$ 1 df
Seattle SNPs p2						
Synonymous	67	102	41			
Nonsynonymous	59	109	45			
% Nonsynonymous	46.8%	51.7%	52.3%			
Significance				G=0.91 P=0.64 2 df	G=0.89 $P<0.35$ 1 df	G=0.01 P=0.92 1 df

[1]Results for the G-test on the entire 2 x 3 table.
[2]Results for the G-test on the 2 x 2 table comparing private vs. shared SNPs.
[3]Results for the G-test on the 2 x 2 table comparing EA private SNPs vs. AA private SNPs.

Table 5.5: Distribution of the Applera SNPs in a sample of 18 chromosomes from each population.

Data set	Count Shared	Count Private AA	Count Private EA	Total Table[1]	Shared vs. Private[2]	AA private vs. EA private[3]
Synonymous	6678.7	7668.3	2884.9			
Nonsynonymous	4640.8	6585.9	3260.5			
% Nonsynonymous	41.0%	46.2%	53.1%			
Significance				G=236.4 $P<4.8\times10^{-52}$ 2 df	G=155.6 $P<1.1\times10^{-35}$ 1 df	G=80.8 $P<2.6\times10^{-19}$ 1 df
Benign	3631.8	4438.1	2052.5			
Possibly damaging	605.3	1266.5	663.7			
Probably damaging	320.2	735.6	469.5			
% Probably damaging	7.0%	11.4%	14.7%			
Significance				G=267.3 $P=1.3\times10^{-56}$ 4 df	G=241.2 $P<4.3\times10^{-53}$ 2 df	G=26.1 $P=2.2\times10^{-6}$ 2 df

[1]Results for the *G*-test on the entire 2 x 3 table.
[2]Results for the *G*-test on the 2 x 2 table comparing private vs. shared SNPs.
[3]Results for the *G*-test on the 2 x 2 table comparing EA private SNPs vs. AA private SNPs.

We hypothesized that the proportional excess of nonsynonymous polymorphism in the EA sample could be due to varying efficacy of purifying selection due to differences in demographic histories between the two populations. Our hypothesis has two testable predictions: 1) if this proportional excess of nonsynonymous polymorphisms in EAs is due to an excess of damaging alleles, we would also expect to find a proportional increase of "probably damaging" SNPs as predicted by PolyPhen in the EA sample, and 2) we should be able to recapitulate this

pattern using simulations with reasonable demographic parameters. When dividing nonsynonymous SNPs into the three PolyPhen categories, we find a significant excess of "probably damaging" SNPs in private SNPs compared to shared SNPs (Table 5.2 and Table 5.3). When considering only the private SNPs, we find a significantly higher proportion of "probably damaging" SNPs in the EA sample relative to the AA sample ($P < 3.3 \times 10^{-11}$, Table 5.2 and Table 5.3), supporting our hypothesis that the excess proportion of nonsynonymous SNPs in the EA sample is due to a higher proportion of damaging SNPs.

In order to assess whether these observations are consistent with plausible demographic histories of the two populations, we developed a large-scale forward simulation program that includes non-stationary demography and a negative log-normal distribution of selective effects for deleterious mutations. Our program used demographic parameters estimated from the data and the literature (Voight *et al.* 2005) for each population (Table 5.6). For example, for the simulations in Figure 5.2a,b, we used a population expansion model for the AAs and a bottleneck model for the EAs (Figure 5.3). We sampled from these simulated populations and found that the proportion of nonsynonymous SNPs is greater in the bottlenecked population than in a population that has expanded (Figure 5.2a; Figure 5.4a; Table 5.6). Furthermore, as shown in Figure 5.2a, the simulated proportions agree with the observed proportions for the Applera dataset (here the proportion includes all SNPs, not just private ones). For all demographic models considered, we observed a higher proportion of nonsynonymous SNPs in the population that underwent a bottleneck as compared to a population of constant size, or that has expanded; the degree to which these other models fit the observed data is variable, however (Table 5.6; Figure 5.4a). For all models tested, we find that a higher proportion of SNPs in the simulated EA

Table 5.6: Description of models used for forward-simulations

Model	Population	Description	$N_{ancestral}$	$T_{ancestral}$	N_{middle}	T_{middle}	N_{last}	T_{last}	ms Code
AA 1	Africa	Constant size	10000	100000					./ms 30 1000 -t 3200
AA 2	Africa	Old growth[1]	7778	100000	25636	6809			./ms 30 1000 -t 8203.52 -eN 0.06640076 0.3034015
AA 3	Africa	Old growth[2]	10000	100000	18000	7500			./ms 30 1000 –t 5760 –eN 0.1041667 0.5555556
EA 1	Europe	Long bottleneck, growth[1]	7895	100000	5699	7703	30030	874	./ms 40 1000 -t 9609.6 -eN 0.007276057 0.1897769 -eN 0.0714036 0.2629038
EA 2	Europe	Short, older bottleneck[3]	10000	100000	1000	800	10000	2400	./ms 40 1000 -t 3200 -eN 0.06 0.1 -eN 0.08 1
EA 3	Europe	Short, recent bottleneck[3]	10000	100000	1000	400	10000	1200	./ms 40 1000 -t 3200 -eN 0.03 0.1 -eN 0.04 1
EA 4	Europe	Short, recent, severe bottleneck[3]	10000	100000	500	400	10000	1200	./ms 40 1000 –t 3200 –eN 0.03 0.05 –eN 0.04 1
EA 5	Europe	Long, less severe bottleneck[3]	10000	100000	3333.3	2400	10000	800	./ms 40 1000 -t 3200 -eN 0.02 0.33333 -eN 0.08 1
EA 6	Europe	Short, recent, severe bottleneck, growth[1,3]	7895	100000	789.5	400	30030	1200	./ms 40 1000 –t 9609.6 –eN 0.00999001 0.02629038 –eN 0.01332001 0.2629038

Notes: $N_{ancestral}$ is the earliest ancestral effective population size. $T_{ancestral}$ is the number of generations simulated using the $N_{ancestral}$ population size. N_{middle} is the effective population size for the second epoch, and T_{middle} is the number of generations simulated in the second epoch. N_{last} is the effective population size in the third epoch, and T_{last} is the number of generations simulated in the third epoch. Finally, the ms commands used to error-check the neutral forward simulations are given. Note, for all simulations, the per locus mutation rate, $\mu = 0.08$. θ is found by multiplying μ by 4 and the appropriate N_e.
[1]Parameters are from Boyko *et al.* (2008); [2]Parameters are from Marth *et al.* (2003); [3]Parameters are from Voight *et al.* (2005).

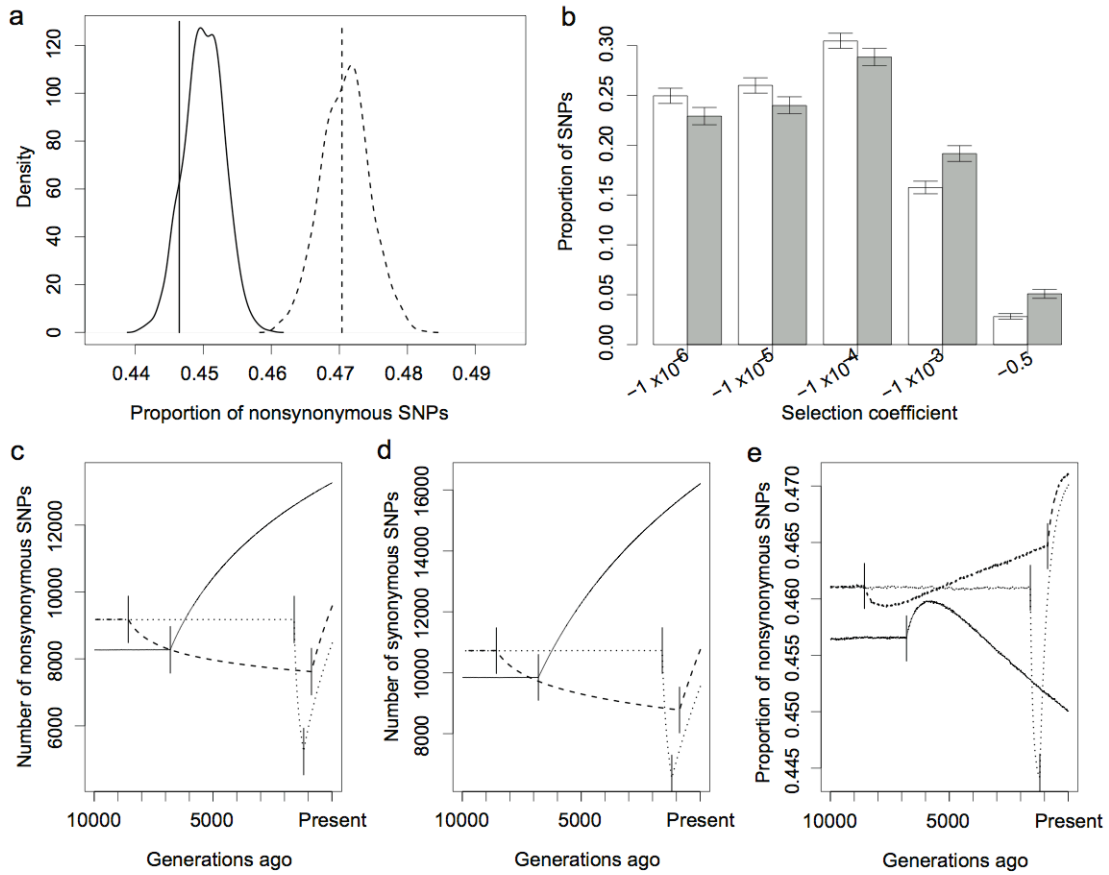


Figure 5.2: Demography and selection can cause a proportional excess of nonsynonymous SNPs in Europeans. **a,b** Results of forward simulations of a population that expanded (AA 2 in Table 5.6), to represent the African American (AA) population and a population that experienced a bottleneck to represent the European (EA) population (EA 1 in Table 5.6). **a,** Distribution of the proportion of nonsynonymous SNPs segregating in samples simulated under European (dashed curve) and African (solid curve) demographic models. Vertical lines show the observed proportions in the Applera dataset. **b,** Distribution of selection coefficients for simulated SNPs in the AA (white bars) and the EA (shaded bars) samples. The labels on the x-axis are the more negative limits of the bins. Error bars denote 95% intervals on the proportion of SNPs in each group. **c-e,** Expected distribution of SNPs over time during a population expansion (AA 2, solid lines), a long, mild bottleneck (EA 1, dashed lines), and a short, severe bottleneck (EA 6, dotted lines). Time moves forward in the figures from left to right. Solid vertical lines indicate when the populations changed size. Further details are given in Table 5.6. **c,** The number of nonsynonymous SNPs, **d,** the number of synonymous SNPs and **e,** the proportion of nonsynonymous SNPs.

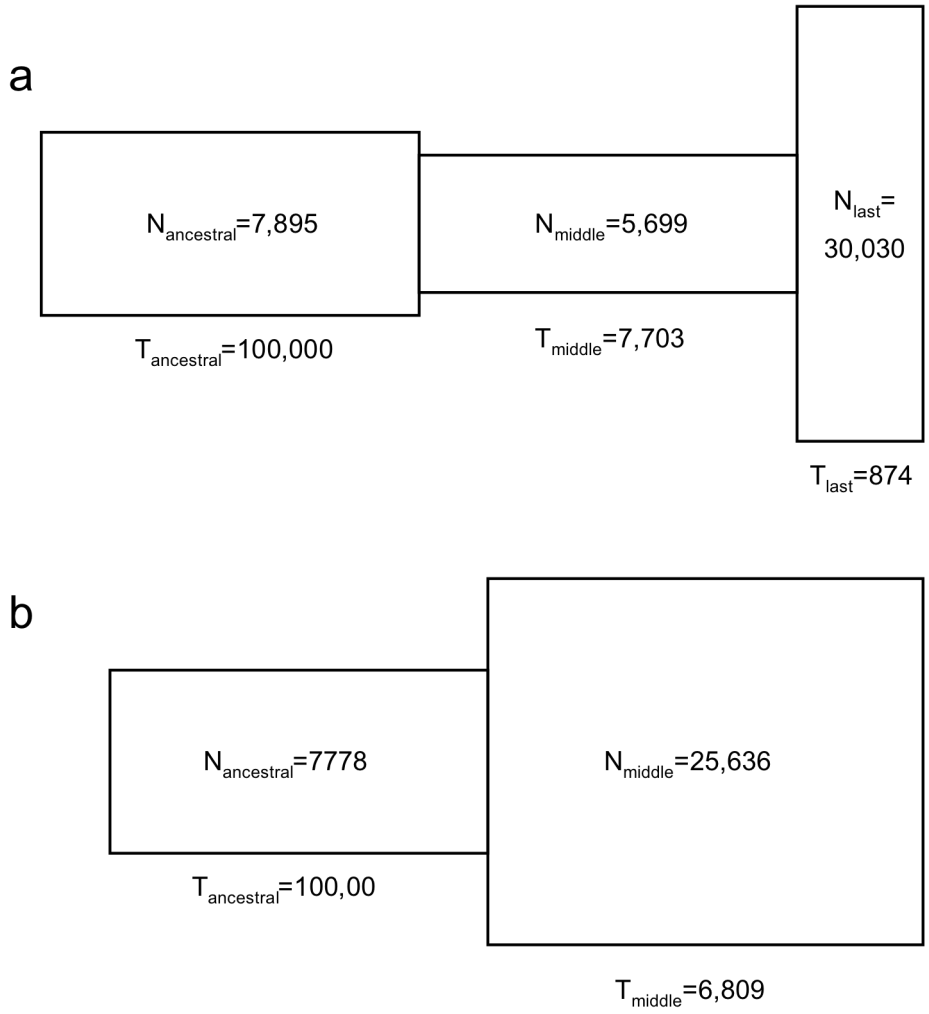


Figure 5.3: Summary of demographic models used for the forward simulations. **a**, Model EA 1. **b**, Model AA 2. $N_{ancestral}$, N_{middle}, N_{last} are the number of diploid individuals in each epoch. Times are in units of generations and moves forward from left to right. Other model parameters are in Table 5.6.

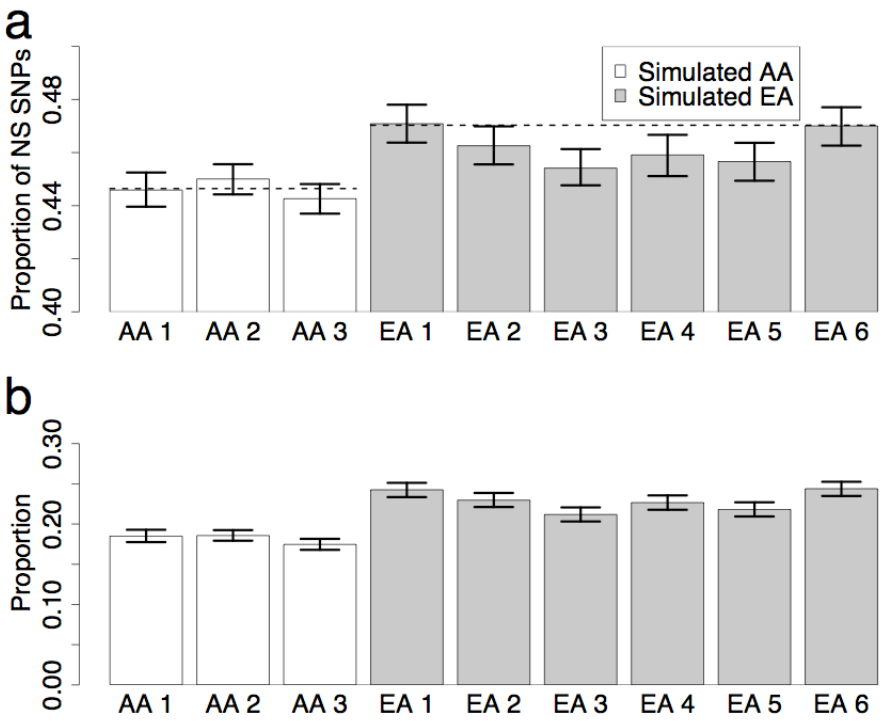


Figure 5.4: Additional results from the forward-simulations with different demographic parameters. The parameters for these models are shown in Table 5.6. The error bars denote 95% intervals obtained from the simulations. **a,** Mean proportion of nonsynonymous SNPs in data simulated for a given demographic model. The horizontal dashed lines indicate the actual proportions of nonsynonymous SNPs observed in the Applera data. Note that the simulated proportion of segregating nonsynonymous SNPs is higher for all the EA models (gray) than for all the AA models (white), consistent with the observed proportions. **b,** Mean proportion of simulated SNPs that are weakly to strongly deleterious ($-1 \times 10^{-3} < s < -0.5$). Note that this proportion is higher for all EA models (gray) than for all AA models (white).

sample are weakly or strongly deleterious ($-0.001 < s < -0.5$) than in the simulated AA sample (Figure 5.2b; Table 5.6, Figure 5.4b), which supports our hypothesis that a higher proportion of deleterious alleles have accumulated in the bottlenecked population. Our analysis illustrates that plausible models of human demography and purifying selection are sufficient to account for the observed increase in the proportion of nonsynonymous SNPs in the EA sample relative to the AA sample.

To determine how the bottleneck contributed to the increased proportion of nonsynonymous SNPs in the EA sample, we recorded the number of SNPs at different time points throughout our forward simulations (see Supplementary Information in APPENDIX 2). Figsure 5.2 c-e show how the number of synonymous SNPs, nonsynonymous SNPs, and the proportion of nonsynonymous SNPs change over time for the EA and AA models described above as well as for a second bottleneck model, having a shorter, but more severe reduction in population size. At the start of the bottleneck, the proportion of nonsynonymous SNPs drops below the pre-bottleneck value (due to the preferential loss of low frequency nonsynonymous SNPs). Then, the proportion increases during the bottleneck due to the accumulation of slightly deleterious SNPs that almost behave neutrally in the small population but are eliminated efficiently from larger populations (Ohta 1973). Once the population expands, the proportion of nonsynonymous SNPs increases dramatically since the increase in population size results in many more mutations (most of which are nonsynonymous, due to the genetic code) entering the population (Figure 5.2c, d). Since growth was recent, purifying selection has not had sufficient time to decrease the proportion of nonsynonymous SNPs to the equilibrium value for the larger population. A related effect has been noted in spatial expansion models, where deleterious mutations can "surf" to high frequency on the edge of the expansion (Travis *et al.* 2007). Our simulations for African demography suggest that once the African population expanded, the proportion of nonsynonymous SNPs also increased initially. But, since the African expansion occurred further back in time than the most recent European expansion, the proportion of nonsynonymous SNPs has had more time to decrease closer to the equilibrium value in the AA sample. At the present time, the absolute numbers of SNPs are higher in the non-bottleneck model (AA 2) than in

the bottleneck models (EA 1 and EA 6). The bottleneck dynamics were robust to the distribution of selective effects used in our simulations (Figure 5.5).

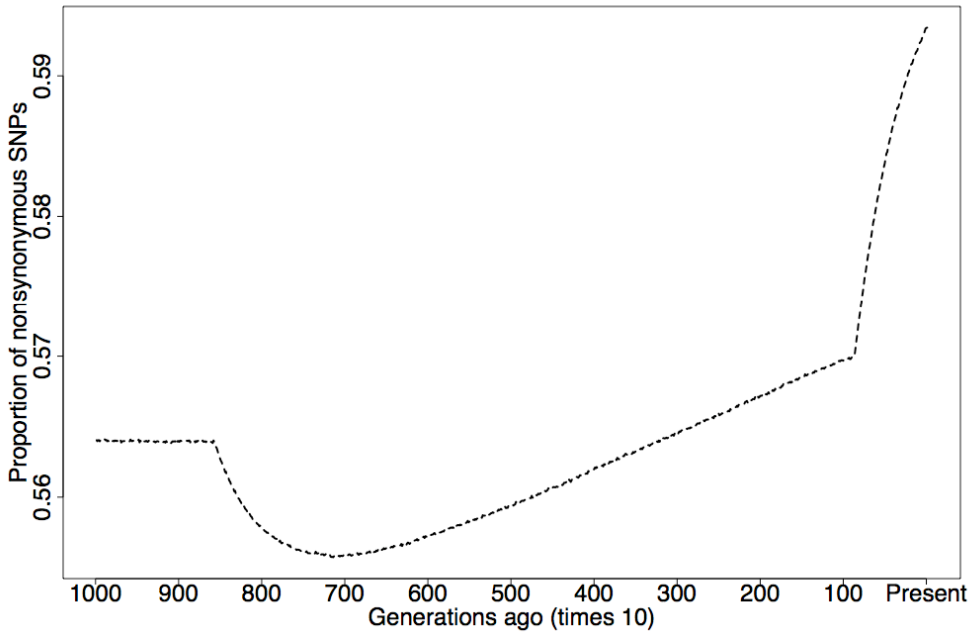


Figure 5.5: Proportion of nonsynonymous SNPs averaged over 1,000 simulation replicates for model EA 1, assuming that all nonsynonymous mutations have a selection coefficient of s = -0.000195. Note that although the proportion of nonsynonymous SNPs is much higher than what we see in the observed data, the proportion changes over time in a similar manner to how it changes when nonsynonymous mutations have selection coefficients drawn from the negative log-normal distribution of selective effects (compare to Figure 5.2e).

Thus, both the PolyPhen analysis and the forward simulations suggest that given the lower levels of genetic diversity compared to Africans, EAs have a higher proportion of deleterious alleles which can be explained by the Out-of-Africa bottleneck and subsequent expansion that outbred European populations endured. This result is important for two reasons. First, while previous work has highlighted examples of European-specific positive selection (Akey *et al.* 2004; Evans *et al.* 2005; International HapMap Consortium 2005; Mekel-Bobrov *et al.* 2005; Voight *et al.* 2006), the importance of adaptations for the evolution of European populations needs to be tempered by our finding that negative selection is less effective at removing

slightly deleterious alleles from European populations. Second, the idea that bottlenecks and founder effects could lead to an increase of damaging alleles in human populations was historically reserved for isolated populations that experienced severe founder effects (*e.g*. Ashkenazi Jews (Slatkin 2004) and Finns (Kere 2001)). Our work suggests that the interaction of demographic processes and purifying selection can have an important impact on the distribution of deleterious variation, even in populations that did not undergo a severe founder effect.

5.3 Methods Summary

We used an improved bioinformatics pipeline to analyze SNPs described in Bustamante *et al.* (2005). We mapped the SNPs to the RefSeq v18 gene model to determine whether they were synonymous or nonsynonymous. Ancestral and derived states for each SNP were determined using the syntenic net alignments between hg18 and panTro2 (Karolchik *et al.* 2003; Kent *et al.* 2003). When counting the number of genotypes per individual, we added a correction for mis-identification of the ancestral allele (Hernandez *et al.* 2007a). SNPs were dropped from the analysis if they failed to meet our bioinformatics quality controls, but we did not filter SNPs based upon frequency.

To predict whether a nonsynonymous SNP will damage protein function, we used an updated version of PolyPhen which has false-positive and false-negative rates below ~15% (Supplementary Methods in APPENDIX 2). When counting the number of damaging genotypes per individual, we used the subset of SNPs where the predicted damaging allele was the derived allele.

An additional four AA individuals were sequenced, but we did not include them (or SNPs private to them) in further analyses since we determined that they had substantially more European admixture than the other AAs (Supplementary Methods

in APPENDIX 2, Figure 5.6, and Table 5.7). If our estimates of admixture are not perfect, this should not drastically affect the comparisons of different classes of SNPs, making our analysis robust to this problem (Supplementary Note 3 in APPENDIX 2). The Coriell sample numbers for the individuals used in our study are given in Table 5.1.

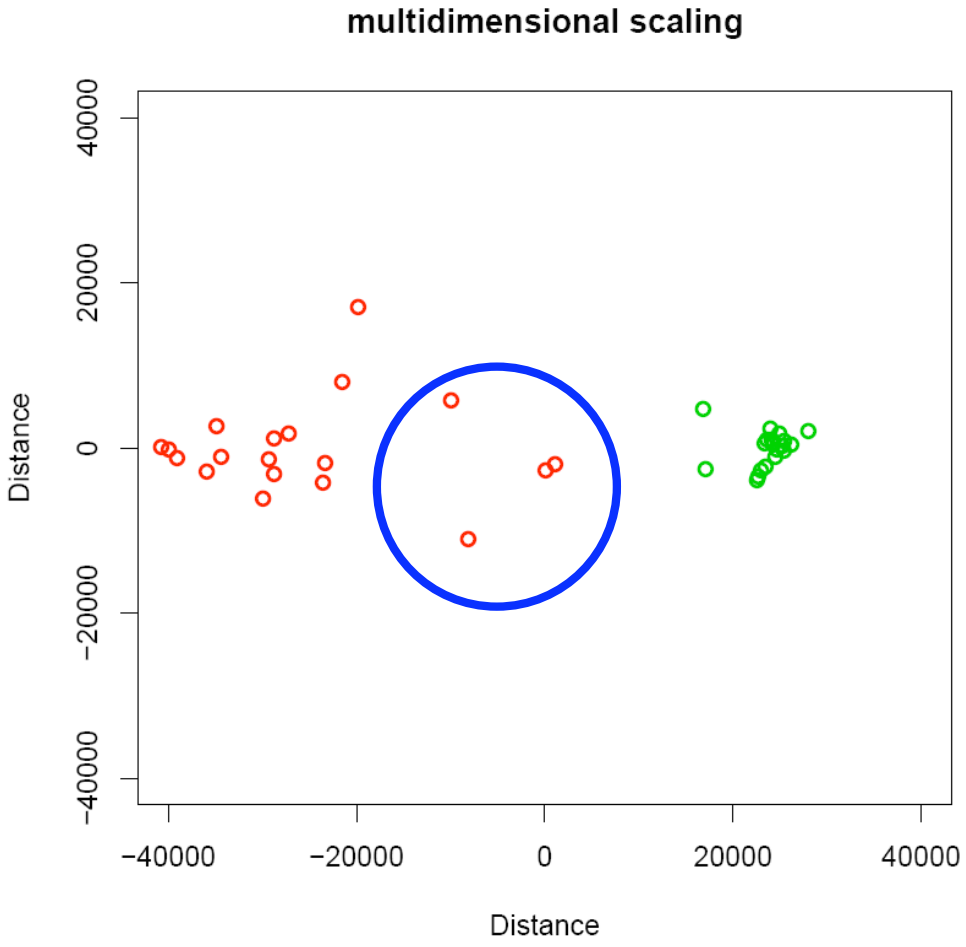


Figure 5.6: Multidimensional scaling based on average genetic identity between individuals. Distances along the first two eigenvectors are shown, with the first eigenvector shown on the x-axis. African-Americans are shown in red and European-Americans are shown in green. The four admixed individuals removed from the analysis are shown inside the blue circle.

Table 5.7: Results of admixture analysis.

Accession	% African ancestry[1]	LR Europe[2]	LR Africa[3]
NA00131	0	0	19400.16772
NA00333	0	0	22911.64034
NA00546	0	0	22385.28786
NA00607	0	0	23205.47219
NA00893	0	0	24303.70296
NA00946	0	0	24176.43834
NA01805	0	0	23268.91208
NA01814	0	0	22878.24185
NA01953	0	0	22685.42073
NA01954	0	0	23215.54464
NA01990	0	0	23203.32778
NA02254	0	0	23820.61363
NA05920	0	0	23465.67788
NA08587	0	0	21945.38032
NA09947	0	0	22211.11724
NA10924	0	0	20792.36264
NA10959	0	0	22874.11692
NA12593	0	0	22783.33802
NA14439	1	38640.75133	0
NA14448	0	0	22579.2436
NA14454	1	34791.20603	0
NA14464	1	43385.57715	0
NA14474	0.698	25363.21432	3682.248693
NA14649	1	41224.16284	0
NA14663	1	37246.60855	0
NA14665	1	40310.99922	0
NA14672	1	42599.04044	0
NA14476	1	40315.4076	0
NA14480	1	40343.18419	0
NA14501	1	44319.97104	0
NA14503	1	37914.1097	0
NA14508	1	40202.15856	0
NA14511	1	38206.12956	0
NA14529	0.597	22949.74517	6328.61658
NA14532	0.601	21718.36213	5761.561953
NA14535	1	41607.0592	0
NA14548	0	0	20580.35807
NA14632	1	42366.3095	0
NA14661	0.807	29920.57209	1677.363973
NA09899	0	0	1119.727854
NA14700	1	14.817113	0

[1]Maximum likelihood estimate of the percentage of African genetic ancestry in each individual.
[2]The –log-likelihood ratio of the maximum likelihood value compared to 100% European ancestry. [3]The –log-likelihood ratio of the maximum likelihood value compared to 100% African ancestry.

To test whether the higher proportion of nonsynonymous SNPs in EAs compared to AAs could be due to the different demographic histories of the two populations, we used forward simulations which allowed us to model demography and purifying selection. We considered a range of demographic models for both populations (Table 5.6) and a distribution of selective effects for nonsynonymous SNPs.

5.4 Online Methods

Bioinformatic pipeline

SNPs were mapped onto RefSeq v18 gene model in a two step process. First we aligned the Celera gene models to hg18 using Blat v33.2 (Kent 2002), filtering out any hits that had less than 98.5% sequence identity or less than 90% coverage. We then aligned RefSeq v18 CDS sequences28 to hg18 using the same filtering conditions. Having coordinates of both our SNPs and RefSeq gene models relative to the assembly, we converted our SNP positions onto the RefSeq CDS position to determine reading frame. If a SNP mapped to multiple RefSeqs, we chose the longest transcript for analysis. Any sequences in RefSeq that were not covered by PCR amplicons were excluded from analysis. SNPs that mapped to multiple RefSeqs that were out-of-frame were discarded. SNPs were polarized by the chimpanzee genome using the syntenic net alignments between hg18 and panTro2 (Karolchik *et al.* 2003; Kent *et al.* 2003). SNPs were dropped from the analysis if they aligned to a non-syntenic region in panTro2, neither human allele matched the panTro2 allele, fewer than nine individuals in either population had a successfully called genotypes, or if we detected a departure from Hardy-Weinberg equilibrium (defined as $P<0.01$) using the exact test of Wigginton *et al.* (2005). SNPs mapping to multiple transcripts were only counted once. We used all SNPs passing bioinformatics quality controls, without

filtering for frequency. Certain analyses were also done excluding singletons and are described in Supplementary Note 2 in APPENDIX 2.

Correction for ancestral mis-identification

Misidentifying the ancestral state of a SNP can lead to miscalculating the proportion of homozygous derived SNPs carried by each individual. We accounted for the probability of ancestral misidentification by adapting the method of Hernandez *et al.* (2007a) to model the number of homozygous SNPs carried by each individual. In this model, the number of homozygous SNPs carried by each individual is considered to be a mixture of sites whose ancestral states were correctly identified using the chimpanzee outgroup and those that were not (two unknown quantities). The corrected number of homozygous derived mutations carried by each individual can then be reconstituted by solving for this unknown quantity as a function of the mixture proportions and observed data. Here, the mixture proportions account for the divergence time between human and chimpanzee using a context-dependent mutation model inferred along the human lineage (Hwang and Green 2004).

PolyPhen analysis

We predicted the functional consequences of SNPs using a newer version of PolyPhen that differs slightly from that described in Sunyaev *et al.* (2001) and Ramensky *et al.* (2002). For SNPs mapping to multiple transcripts, we ran PolyPhen on the SNP in each transcript. If a SNP had different PolyPhen predictions in different transcripts, it was excluded from any further PolyPhen analyses. 340 SNPs had multiple PolyPhen predictions and 56 did not have a prediction. For our data, PolyPhen used an average of 18.2 (SD: 28.0) sequences across covered SNPs. SNPs used for analyses, along with their frequencies and PolyPhen predictions are available. For approximately 83.9% of the "benign", 98.2% of the "possibly damaging" and 98.8% of "probably damaging" SNPs, the damaging allele (the allele with the lower

PSIC score) is the derived allele, indicating that PolyPhen has a greater ability to distinguish which allele is damaging for "probably damaging" SNPs than for "benign" or "possibly damaging" SNPs. As explained in the Supplementary Methods in APPENDIX 2, PolyPhen classified 85.5% of 3,604 disease mutations annotated in the UniProt database as either probably or possibly damaging, while predicting 86.1% of 12,237 amino acid differences between humans and another mammalian ortholog as benign. These results suggest that the false positive and false negative rates of the algorithm are each below ~15%.

Counting the number of genotypes per individual

To determine whether AA individuals were heterozygous at more SNPs than EA individuals, we used a two-sided Mann-Whitney U test (MWU) to compare the distribution of the number of heterozygous genotypes per individual in AA individuals to the distribution of the number of heterozygous genotypes per individuals in the EA individuals. This comparison was done separately for synonymous, nonsynonymous, "benign", "possibly" and "probably damaging" SNPs. A similar test was used to test whether EA individuals were homozygous for the derived allele at a greater number of SNPs than EA individuals. When counting the number of SNPs per individual, we wanted to ensure that our counts were not biased because some samples had more complete sequencing than others. We divided the number of genotypes in an individual of each particular category (e.g. number of heterozygous genotypes for synonymous sites in a particular individual) by the total number of genotypes in that category (e.g. total number of genotypes at synonymous sites) in the individual. We then tested if the distribution of these proportions was different between the AA and EA sample. In all cases, we observed the same pattern as shown in Figure 5.1 (data not shown), indicating that this result was not due to inconsistent sequencing of different individuals.

Forward simulations

A detailed description of the methods used for forward simulations is given in Supplementary Methods in APPENDIX 2. Briefly, we wanted to test whether the observation of a higher ratio of nonsynonymous to synonymous SNPs in EAs than in Africans could have been due to the different demographic histories of the two populations. We simulated one population forward in time with a demographic history consistent with that of Africans and another population forward in time with demographic history consistent with that of Western Europe. We considered a variety of plausible demographic models for each population (Voight *et al.* 2005), and simulated the African and European populations independently of each other. In addition to simulating populations where all SNPs were neutral, we also independently simulated a second set of populations for each set of demographic parameters where the selection coefficients were from a distribution of selective effects (Supplementary Methods in APPENDIX 2) to mimic nonsynonymous sites. At the end of the simulation, we sampled 15 individuals from the population that expanded and 20 individuals from the population that underwent a bottleneck. We examined whether one population had a higher proportion of damaging (nonsynonymous) SNPs and whether segregating SNPs in one population had a different distribution of selection coefficient than SNPs segregating in the other population.

APPENDIX 1

SUPPLEMENTARY TEXT AND TABLE FOR CHAPTER 1

Supplementary Text

Haplotype phase uncertainty

Since the *HCN* statistic reflects haplotype patterns, and for many genome-wide SNP datasets consisting of unrelated individuals, haplotype phase would need to be computationally inferred, we wanted to determine how this inference affected the *HCN* statistic. To do this, we simulated 1000 windows with c_{window} = 0.25 cM in a sample size of 100 chromosomes from a bottleneck demographic history (N_{cur}=10,000,N_{mid}/N_{cur}=0.1,N_{anc}/N_{cur}=1.0, t_{cur}=800 generations, t_{mid}=800 generations), where n_{snp}=40. For each window, we then randomly paired the chromosomes into diploid individuals and generated diploid genotypes at each SNP. We next inferred haplotypes from these genotypes using a popular phasing method, fastPHASE (Scheet and Stephens 2006), with the default settings. We chose to use fastPHASE since its performance is comparable to one of the better performing phasing algorithms, PHASE, yet is fast enough to be run on genome-wide datasets. Finally, we compared the *HCN* statistic for the phase-known dataset to the phase-inferred dataset.

Figure 1.8 shows the *HCN* statistic for a bottleneck model when the correct haplotype phase is known with certainty (left) and when haplotype phase is inferred using fastPHASE (right). The *HCN* from phase-inferred haplotypes has a broader distribution than when haplotype phase is known. In particular the *HCN* constructed using the phase-inferred haplotypes has an excess of windows having many haplotypes (green squares in bins "65-90" and "70-100") as compared to the known phase *HCN*. Although it is a bit more subtle, the *HCN* using the phase-inferred haplotypes also has an excess of windows where the most common haplotype is at a

high frequency. This can be seen by the yellow square in the phase inferred haplotypes where there was an orange square in the phase-known *HCN*. Thus, inferring haplotype phase will result in an *HCN* statistic that is slightly different from the true phase-known *HCN*.

Ascertainment bias

To evaluate how the *HCN* statistic is influenced by SNP ascertainment bias, we conducted a variety of coalescent simulations under different demographic models and SNP ascertainment strategies. We then compared the *HCN* from the different ascertainment strategies to the *HCN* with complete SNP ascertainment. We also examined whether another haplotype statistic, H_{pair}, is affected by ascertainment bias.

Since we wanted to address the question of whether discovering SNPs in one population and then typing them in a second population is more biased than selecting the SNPs in the genotyped population, we considered demographic models that consisted of two populations. Briefly, we considered a finite island model (where each population has size N_e=10,000) with a low rate of migration between populations ($4N_em$=9) and high rate ($4N_em$ =99), a population split model where the two populations (each of size N_e =10,000) split 2000 or 5000 generations ago, and a complex model where the two populations split 5000 generations ago and there was a bottleneck in population one (N_{mid}/N_{cur}=0.1,t_{cur}=800, t_{mid}=800). The last model can be thought of as a very crude approximation of the contrast between European (as population 1) and African (as population 2) human populations. For each of these demographic models, we simulated a "genotype" sample of 40 chromosomes from each of the two populations as well as a SNP discovery sample consisting of an additional four chromosomes from each population. We then examined five different SNP discovery protocols shown in Supplementary Table 1A. These ascertainment strategies are reasonable ones for many of the human genome-wide SNP datasets like

HapMap where many of the SNPs were discovered by comparing two sequencing reads (as in phase I) or from a polymorphism discovery panel with a few chromosomes from multiple populations (phase II SNPs discovered by Perlegen) (Hinds *et al.* 2005; International HapMap Consortium 2005; International HapMap Consortium 2007). For each ascertainment scheme we simulated 1000 windows 500 kb in size with a uniform recombination rate of 1 cM/Mb (c_{window}=0.5 cM) and μ=1 x 10^{-8} per base-pair per generation. To determine whether ascertainment bias becomes a problem for larger datasets containing more than 1000 windows (n_{window}>1000), we also simulated an additional dataset under the complex demographic history consisting of 7000 windows 250 kb in size with uniform recombination rate of 1cM/Mb (c_{window}=0.25 cM) and μ=1 x 10^{-8} per base-pair per generation. Finally, we considered the case where the genotype sample consisted of 120 chromosomes from each population (to mimic the HapMap CEU and YRI samples) and we had data from 7000 windows 250 kb in size with uniform recombination rate of 1cM/Mb (c_{window}=0.25 cM) and μ=1 x 10^{-8} per base-pair per generation. For this set of simulations, the SNP discovery set consisted of 12 chromosomes per population. Here we considered eight ascertainment strategies shown in Supplementary Table 1B.

For each demographic scenario and ascertainment scheme, we selected a sub-set of 40 SNPs having MAF >10% (n_{snp}=40). If a window had fewer than 40 SNPs, it was dropped from the analysis. We then generated the *HCN* statistic (again averaging over 10 sets of randomly selected SNPs for each window) and compared the statistic to the expected statistic under complete ascertainment using a chi-square goodness of fit test.

Supplementary Table 1: Summary of SNP ascertainment strategies.

Abbreviation	Ascertainment sample description
A. Figures 1.2, 1.9 &1.10; n=40	
2 from pop 1	2 chromosomes from population 1
1 from each	1 chromosome from population 1 and 1 chromosome from population 2
4 from pop 1	4 chromosomes from population 1
2 from each	2 chromosomes from population 1 and 2 chromosomes from population 2
4 from each	4 chromosomes from population 1 and 4 chromosomes from population 2
B. Figure 1.11; n=120	
2 from pop 1	2 chromosomes from population 1
1 from each	1 chromosome from population 1 and 1 chromosome from population 2
4 from pop 1	4 chromosomes from population 1
2 from each	2 chromosomes from population 1 and 2 chromosomes from population 2
4 from each	4 chromosomes from population 1 and 4 chromosomes from population 2
12 from pop 1	12 chromosomes from population 1
12 from pop 2	12 chromosomes from population 2
12 from each	12 chromosomes from population 1 and 12 chromosomes from population 2

To generate the expected *HCN* statistic under complete ascertainment, we simulated an additional 10^5 windows each consisting of 40 chromosomes and n_{snp}=40 and averaged over 10 sets of randomly selected SNPs for each window. From these simulations, we computed the expected *HCN* statistic. Note that, when conducting the chi-square goodness of fit tests, we binned the *HCN* statistic so that we did not have any expected cell counts ≤ 5. For the complex demographic history using 7000 windows (for genotype sample sizes of 40 and 120 chromosomes per population) n_{snp}=20 instead of 40.

We find that for all demographic models examined, except for the island model with a high migration rate, ascertainment of SNPs using two discovery chromosomes from one population results in a different *HCN* statistic than that expected under complete ascertainment (Figure 1.9). This is shown by the low *P*-values for the goodness of fit tests comparing the *HCN* statistic using SNPs polymorphic in two discovery chromosomes to the expected *HCN* under complete ascertainment. The *HCN* statistic constructed from SNPs ascertained in two chromosomes has an excess of windows having a small number of haplotypes and an excess of windows where the most common haplotype is at higher frequency as compared to the complete ascertainment case (Figure 1.10).

The reason for this pattern is that SNPs polymorphic in the two chromosome discovery sample must occur on branches of the genealogy where one of the two discovery chromosomes carries the mutant allele and the other does not. These branches are a small fraction of the total area of the genealogy. This fact will result in SNPs that are polymorphic in the discovery sample tending to occur on the same branches of the genealogy more often than expected without ascertainment bias. SNPs that co-occur on the same branches of the genealogy will be in LD with each other, resulting in there being fewer haplotypes and the most common haplotype occurring at higher frequency than in the case of less LD among SNPs. When considering SNPs discovered from two chromosomes from the first population, the *HCN*s in both populations differ from the expected *HCN*, suggesting that SNP discovery using two chromosomes does poorly, regardless of whether those two chromosomes are from the population of interest.

SNP discovery using one chromosome from each population is a slight improvement to SNP discovery using two SNPs from population 1. However, we note that for many of the demographic models considered here (Figure 1.9), the *HCN*

constructed from ascertained SNPs differs significantly from the expected *HCN* under complete ascertainment.

However, SNP discovery using four chromosomes from the first population results in a better fit to the expected *HCN* for most of the demographic models considered. In all cases, except for the complex demographic model, the *HCN*s constructed from ascertained SNPs are quite consistent with the expected *HCN* under complete SNP ascertainment. This finding holds true even for the second population which had no SNP discovery, again illustrating that if the two populations have similar demographic histories, ascertainment sample depth may be more important than which population the SNPs were ascertained from in terms of matching the *HCN* statistic. This pattern, however, does not hold for the complex demographic model. Here SNP discovery using four SNPs from the bottlenecked population (population 1) results in a poor fit to the expected *HCN* statistic. The reason for this is that the four SNP discovery chromosomes from the bottlenecked population are less representative of the diversity in the second population that did not undergo a bottleneck (population 2). If, again for the complex demographic scenario, instead of taking four discovery chromosomes from the first population, we take two discovery chromosomes from each population, the *HCN* statistic from the ascertained SNPs more closely matches the expected *HCN* statistic. However, note that if the number of windows of the genome considered is large (n_{window}=7000), the effects of ascertainment bias are still present.

The *HCN* statistic generated using a four chromosome SNP discovery sample from both of the two populations results in an excellent fit to the expected *HCN* for both populations in all demographic scenarios considered. We also found an adequate fit of the expected *HCN* to the observed *HCN* when considering a larger dataset under the complex demographic model. This finding is especially encouraging since the

larger number of windows in the dataset ($n_{winidow}$=7000 as compared to 1000 in previous datasets) will have more power to detect subtle departures in the fit of the model. Thus, for the demographic models considered here using n=40 chromosomes, the *HCN* statistic using SNP discovery sample of ≥4 chromosomes from at least two populations is not significantly different from the true *HCN* statistic.

We also examined whether ascertainment bias is a more severe problem when the genotype sample is >40 chromosomes. To do this, we repeated the above approach for the complex demographic model using n=120 chromosomes and considering larger SNP discovery sample sizes (Figure 1.11). We find that small SNP discovery sample sizes (here <8 chromosomes) result in significant differences between the *HCN* under SNP ascertainment and the expected *HCN*. However, for larger SNP discovery sample sizes, the effect disappears. This holds true even for the population that has no SNP discovery chromosomes (*e.g.* the solid line at "12 from 2"). To assess the amount of evolutionary variance in the whole process, we performed two completely independent sets of simulations for these demographic and ascertainment models. The results of both replicates are shown in Figure 1.11. Encouragingly, the variance is reasonably low since the two solid curves (and dotted curves) are similar to each other.

We also evaluated whether the H_{pair} statistic was robust to ascertainment bias. As shown in Figure 1.2, for all demographic models and ascertainment conditions considered, H_{pair} was severely affected by SNP ascertainment bias. Ascertainment bias results in H_{pair} being higher than expected. This finding is analogous to the effect of ascertainment bias on π, the average number of pairwise differences among DNA sequences (Nielsen *et al.* 2004). Ascertainment bias results in an excess of intermediate-frequency SNPs, which results in there being more pairwise differences

between haplotypes than low-frequency SNPs do. Thus, by preferentially selecting intermediate-frequency SNPs, H_{pair} becomes inflated.

Interestingly, we find that for the cases where SNPs were ascertained in population 1 exclusively, the fit of the H_{pair} statistic under ascertainment bias to the expected H_{pair} statistic is actually worse in population 1—the population where the SNPs were discovered in—than in population 2. This pattern is seen for both n_{widnow}=1000 and for n_{widnow} =7000 and for both the "2 from pop 1" and the "four from pop 1" ascertainment strategies. One possible explanation for this counter-intuitive pattern is that the ascertained SNPs from population 1 are more likely to be at intermediate frequency in population 1 (as discussed above), but may have drifted to lower or higher frequency in the second population, resulting in those SNPs being more representative of the true frequency spectrum in that population.

The ms commands to generate the HCN statistic in Figure 1.18

Growth and Structure

./ms 40 10000 -t 400 -r 400 250000 -F 4 -es 0.00625 1 0.1 -eM 0.00625 5 -eN 0.00625 0.5 -eN 0.025 0.125 -ej 0.625 2 1 -eM 0.625 0 -eN 0.625 0.25

Growth

./ms 40 10000 -t 400 -r 400 250000 -F 4 -en 0.01925 1 0.303333

APPENDIX 2

SUPPLEMENTARY TEXT FOR CHAPTER 5

Supplementary Methods

Estimation of admixture proportions

An important aspect of our analysis is identifying and removing individuals from the data set that exhibit a high proportion of recent admixture of European and African ancestry. To identify such individuals, we used a maximum likelihood approach similar to the Bayesian approach implemented in Structure (Pritchard *et al.* 2000). Briefly, the proportion of an individuals (*i*) genome that is from population *k* is $q_k^{(i)}$, and p_{klj} is the frequency of allele *j* in locus *l* in population *k*. The probability that individual *i* has genotype *jv* in locus *l* is then

$$\Pr\left(x_l^{(i)} = (j,v)\right) = 2\left(\sum_k p_{klj} q_k^{(i)}\right)\left(\sum_k p_{klv} q_k^{(i)}\right), \qquad \text{(S1)}$$

Notice that there is an independence assumption between the two gene copies in each locus. The likelihood function is obtained assuming independence among loci and individuals by multiplying the value of $\Pr\left(x_l^{(i)} = (j,v)\right)$ among all values of *i* and *l*. The likelihood function can be calculated quite easily when individuals can be assigned to populations *a priori*. Optimization of the likelihood function can be done using any standard algorithm (e.g. Press *et al.* 1992 pp. 420).

Computationally this approach is much faster than the Markov chain Monte Carlo (MCMC) method used in Structure (Pritchard *et al.* 2000). The results of the analysis are shown in Table 5.7. Notice that all but four individuals have estimates of the proportion of African ancestry equal to either zero or one. All four admixed individuals were self-reported as African-American. We emphasize that with the size

of this dataset, allele frequencies may not be estimated well, which could create biases in the inferences of admixture parameters. In addition, unrealistic assumptions regarding independence among individuals and among SNPs may affect our results. However, cluster analysis (Figure 5.6) corroborates our results, that with the exception of the four individuals, individuals cluster in two discrete populations. These analyses justify using the remaining individuals as approximations of original European and African populations. However, we emphasize that for more detailed inferences of human demography it is desirable to use individuals of with known origin within Africa.

Forward simulations

We used forward simulations to determine whether observed patterns of amino acid and silent variation could have been accounted for by the interaction of demography and negative selection. Since the precise demographic history of humans in unknown, we considered different sets of parameters consistent with what has been reported in the literature. Table 5.6 shows the different sets of parameters that we considered. For the African population, we considered a model of constant population size (AA 1) and two models with ancient population growth (AA 2 and AA 3). For Europeans, we used a model with a very long, recent bottleneck that was not severe (EA 1), a short, but older, bottleneck (EA 2), short, recent, bottlenecks of varying severity (EA 3, EA 4, and EA 6) as well as a long, recent, but not severe bottleneck (EA 5).

For each of the nine models in Table 5.6, we conducted two sets of forward simulations: 1) where all mutations were neutral and 2) where the selection coefficient for a particular mutation came a distribution of selective effects. The neutral class was to represent synonymous sites while the selected class was to represent nonsynonymous sites. The nonsynonymous sites are best represented by a distribution

of selective effects since some the fitness consequences of novel amino acid mutations are known to vary widely. The log-normal distribution of selective effects we use captures this pattern (see below).

We ran 1,000 replicates of the simulation for each of the nine demographic models, both with and without selection. The African and European populations were simulated independently of each other and modeled as Wright-Fisher panmictic populations. For each replicate, a Poisson number of mutations with mean $\theta/2$ entered each generation, where $\theta = 4N_e\mu$, N_e is the effective diploid population size and μ is the per locus mutation rate. For all simulations, we used a constant $\mu = 0.08$ for silent sites and changed N_e as reflected in Table 5.6. Thus, θ (and as a result, the number of mutations entering every generation) was appropriately scaled by the effective population size. To model nonsynonymous SNPs, we set $\mu = 0.2$, to reflect the fact that we expect the nonsynonymous mutation rate to be 2.5 times higher than the synonymous mutation rate. All mutations were assumed to be independent, and we chose μ so as to match approximately the total observed number of silent SNPs observed in the data. The frequencies of segregating mutations were changed by random binomial sampling to mimic genetic drift. The simulation was run for an initial burn-in period of 100,000 generations to achieve stationarity. The population size change in all models (expect for the constant size model) was done by simply changing the number of alleles to be drawn during the binomial sampling step to match the new population size. After running the simulation for the appropriate number of generations, we took samples of 30 alleles for models of African demography or 40 alleles for models of European demography.

For the simulations using selection, we used a log-normal distribution of selective effects. Based on our other work (Boyko *et al.* 2008) we have found that a negative log-normal distribution for γ (the scaled selection coefficient where $\gamma=2N_es$)

with a mean of γ= 5.02 and standard deviation of 5.94 provides an excellent fit to the frequency distribution of nonsynonymous SNPs. We converted γ into the selection coefficient, *s*, by dividing γ by 51,272, or twice the estimated current African effective population size estimated from our dataset. For each mutation in the selected class that entered the population, we obtained a selection coefficient, *s*, by drawing from the log-normal distribution above. Since *s* (as scaled here) is bounded between -0.5 and 0.5, if we obtained an s <-0.5, we simply set it equal to -0.5. Each generation we adjusted the frequency of the mutations both by binomial sampling (genetic drift) as well as deterministically using the standard selection equation (e.g. Hartl and Clark 2007). In our simulation, the fitness of the homozygote not carrying the mutation of interest is 1, the fitness of the heterozygote is 1+*s*, and the fitness of the other homozygote is 1+2*s*. Thus, all mutations have an additive effect on fitness, which was the model assumed for identifying the distribution of selective effects (Boyko *et al.* 2008). Full details of the inference of the distribution of selective effects on new mutations will be presented in Boyko *et al.* (2008).

All simulations were written in C and were run on a 101 node Apple G5 cluster. Random numbers were generated using the GNU Scientific Library (http://www.gnu.org/software/gsl/). To error check the code for the neutral simulations, we compared the site frequency spectrum obtained from the forward simulations to the site frequency spectrum simulated using coalescent theory as implemented in the computer program ms (Hudson 2002). The ms commands used are given in Table 5.6.

To examine how the proportion of nonsynonymous SNPs changes over time, we used a similar forward-simulation scheme as described above, this time keeping track of the proportion of nonsynonymous SNPs over time. For computational efficiency, we only considered models EA 1, EA 6, and AA 1 (see Table 5.6). When

running the simulations, we took a sample of 15 individuals for model AA 1 and 20 individuals for models EA 1 and EA 6 every ten generations, starting 10,000 generations ago. We then recorded the proportion of nonsynonymous SNPs segregating in these samples. Figure 5.2e shows how the proportion (averaged over 1,000 simulation replicates) changes over time. To determine what effect the log-normal distribution of selective effects had on the dynamics of the proportion of nonsynonymous SNPs, we repeated the simulation using model EA 1, where this time, all nonsynonymous mutations had the same selection coefficient, s = -0.000195. While the ending proportion of nonsynonymous SNPs is much higher than what we see using the distribution of selective effects, the change in proportion over time follows a similar shape as what we see when using a distribution of selective effects. This can be seen by comparing the curve for model EA 1 in Figure 5.2e to the curve shown in Figure 5.5.

Projection of allele frequency to smaller sample size

To determine what effect using different sample sizes for the AA (15 individuals) and EA samples (20 individuals) and to what degree differences in missing data influenced our analysis of the differences in the proportions of SNPs between the AA and EA samples, we repeated our analysis taking these factors into account. More specifically, we projected our data down to a smaller sample size of 9 individuals (18 chromosomes), since this was the maximum amount of missing data we allowed each SNP. Doing this ensured that all SNPs had the exact same sample size and amount of missing data in both populations. Thus, any differences in the distribution of SNPs in this analysis cannot be due to sampling differences between populations. Before describing this analysis, we will introduce some notation. For each SNP in our dataset, let N_1 denote the total number of chromosomes surveyed at this SNP in the EA sample (N_1 ranges from 18 to 40), N_2 denote the total number of

chromosomes surveyed at this SNP in the AA sample (N_2 ranges from 18 to 30), X_1 denote the number of chromosomes carrying the derived allele in the EA sample (X_1 ranges from 0 to N_1), and X_2 denote the number of chromosomes carrying the derived allele in the AA sample (X_2 ranges from 0 to N_2). While X_1 and X_2 represent the derived allele counts, this analysis does not need to distinguish between the ancestral and derived state, and could easily be replaced by picking one allele arbitrarily, as long as X_1 and X_2 represent the counts of the same allele.

We can compute the probability that a SNP at frequency X_1 in a sample of size N_1 would have a different frequency, say Y_1, in a sample of size 18 using the hypergeometric distribution. For example we can calculate the probability that a SNP has a frequency of 0 in the smaller sample of size 18, as

$$P(Y_1 = 0) = \frac{\binom{X_1}{0}\binom{N_1 - X_1}{18}}{\binom{N_1}{18}}, \qquad \text{(S2)}$$

The probability that the SNP would have frequency 18 out of 18 in the new EA sample is computed as

$$P(Y_1 = 18) = \frac{\binom{X_1}{18}\binom{N_1 - X_1}{0}}{\binom{N_1}{18}}, \qquad \text{(S3)}$$

The exact same calculation is made for the AA sample by replacing Y_1 with Y_2, X_1 with X_2, and N_1 with N_2.

Now that we have computed $P(Y_1 = 0),\ P(Y_1 = 18),\ P(Y_2 = 0),\ P(Y_2 = 18)$ as described above, we can find the probability that the SNP is monomorphic, private to EA, private to AA, or shared between the two populations. The probability that a SNP is monomorphic can be written as

$P(monomorphic) = P((Y_1 = 0 \cup Y_1 = 18) \cap (Y_2 = 0 \cup Y_2 = 18))$, which can easily be calculated by $P(monomorphic) = (P(Y_1 = 0) + P(Y_1 = 18)) * (P(Y_2 = 0) + P(Y_2 = 18))$. Note, if the new samples were to become fixed for opposite alleles, that is, $Y_1 = 0$ and $Y_2 = 18$ or $Y_1 = 18$ and $Y_2 = 0$, we still considered that SNP to be monomorphic, since it would not easily fit into the shared, private AA, or private EA categories used for our analysis. The probability that a SNP is private to the EA sample can be written as $P(private\ EA) = P((Y_2 = 0 \cup Y_2 = 18) \cap (Y_1 \neq 0 \cap Y_1 \neq 18))$, and can be calculated by $P(private\ EA) = P(Y_2 = 0) + P(Y_2 = 18) - P(monomorphic)$. Similarly, the probability that a SNP is private to the AA sample can be written as $P(private\ AA) = P((Y_1 = 0 \cup Y_1 = 18) \cap (Y_2 \neq 0 \cap Y_2 \neq 18))$, and can be calculated by $P(private\ AA) = P(Y_1 = 0) + P(Y_1 = 18) - P(monomorphic)$. Finally, the probability that a SNP is shared between the two populations can be calculated as $P(shared) = 1 - P(monomorphic) - P(private\ AA) - P(private\ EA)$.

The above computations were carried out for each SNP and summed together, giving the number of SNPs in each of the four categories for the new sample size of 18 for both the EA and AA sample. We carried out the above calculations separately for synonymous and nonsynonymous SNPs as well as the three PolyPhen categories. The results of these calculations are shown in Table 5.5.

Performance of PolyPhen

To estimate the sensitivity and specificity for the version of PolyPhen employed in this manuscript we used a set of 3,604 disease mutations annotated in the UniProt database. We only selected nonsynonymous mutations with a clear description of the effect of mutations on phenotype. We also used a set of 2,603 mutations derived from site-directed mutagenesis experiments with known effect on protein function. PolyPhen correctly predicted 80.1% and 85.5% of disease and mutagenesis mutations respectively as damaging (53% and 68.9% as probably

damaging and 27.1% and 16.6% as possibly damaging). We note that estimates of sensitivity are higher if more stringent datasets with clearer description of functional and phenotypic effect of mutations are used. We used 12,237 substitutions between the same human proteins and closely related mammalian orthologs to estimate specificity (true negative rate). In this set 86.2% were correctly predicted as benign, 5.5% were incorrectly predicted as probably damaging and 8.4% as possibly damaging. It is difficult to estimate false-discovery rate because the true fraction of completely benign amino acid changes among human polymorphism is unknown. Approximately 28% of nonsynonymous SNPs in the Applera dataset are likely damaging to protein function. This estimate is very similar to previous studies (Fay *et al.* 2001; Sunyaev *et al.* 2001; Ng and Henikoff 2006; Yue and Moult 2006) using SNPs from dbSNP.

Supplementary Note 1

Analysis of Seattle SNPs data

To ensure that the proportional excess of nonsynonymous SNPs in Europeans (EAs) relative to African Americans (AAs) in the Applera dataset was not an artifact of the sequencing or SNP calling pipeline, we repeated our analyses using data obtained by the Seattle SNPs project. They sequenced genes in two different panels of individuals. 211 genes were sequenced in the first panel (herein p1) consisting of 24 AAs and 23 EAs. 76 genes were sequenced in the second panel (herein p2) consisting of 24 Yoruba individuals and 23 CEPH individuals. The individuals in p2 are a subset of those genotyped by the International HapMap Consortium. We downloaded all data for 287 genes on December 3, 2006.

We determined the number of synonymous and nonsynonymous SNPs in both these panels. The 2 x 3 tables displaying these data are shown in Table 5.4. For panel p1, we find a significant increase in the proportion of private nonsynonymous SNPs in

the EA sample (0.667) relative to the AA sample (0.510; $G = 13.0$, $P < 3.1 \times 10^{-4}$, 1 df), supporting what we found in the original analysis of the Applera data. As expected, we also found that the proportion of nonsynonymous SNPs is higher for private SNPs (0.552) than for SNPs shared between the two populations (0.484; $G = 4.5$, $P < 0.035$, 1 df). However, for panel p2, we did not find a significant increase in the proportion of private nonsynonymous SNPs in the EA sample compared to the Yoruba sample ($G = 0.01$, $P = 0.92$,1 df), nor did we find a significant increase in the proportion of nonsynonymous private SNPs compared to nonsynonymous shared SNPs ($G = 0.89$, $P < 0.35$, 1 df).

We hypothesized that the reason we did not find a significant increase in the proportion of private nonsynonymous SNPs in the EA sample compared to the Yoruba sample (p2) because the number of SNPs included in this dataset was quite small. To determine the power of this analysis, we performed Monte Carlo simulations assuming the effect sizes observed in the Applera dataset are the true effect sizes. We estimated that about 47.0% of private SNPs in the Applera AA sample were nonsynonymous and 55.4% of private SNPs in the EA sample were nonsynonymous (Table 5.2). Additionally, as shown in Table 5.4, there were 86 private SNPs in the p2 CEPH (EA) sample and 211 private SNPs in the p2 Yoruba sample. We simulated 10,000 datasets by taking 86 and 211 draws from binomial distributions with probability of success equal to the proportion of nonsynonymous SNPs (0.554 and 0.470, respectively). We then performed the *G*-test with 1 df on these simulated datasets, and the proportion of sets where $P < 0.05$ is the power of the analysis. Under these assumptions, our analysis had 26.14% power to detect this effect. Conversely, when we use a similar method to estimate the power of the p1 analysis (using 484 SNPs in the AA sample and 177 SNPs in the EA sample), we found that this analysis has 49.25% power. Thus, given the sample sizes of the two different datasets from SeattleSNPs, it is not

unexpected that we found different results. While the analysis of the p2 data has low power due to the small number of SNPs discovered, we cannot rule out that there are truly different effects occurring in the Yoruba than in African Americans, though additional analyses using HapMap data suggest otherwise (Supplementary Note 3 in APPENDIX 2).

Supplementary Note 2

Quality control of the Applera data

In this section we describe the analyses that were done to ensure that the proportional excess of nonsynonymous SNPs in the EA sample relative to the AA sample in the Applera data is not due to sequencing error. Here we will describe two types of analyses that support this assertion. 1) We will present a modified analysis excluding SNPs that could be due to sequencing artifacts, and 2) we show that an unreasonably high sequencing error rate would be required to explain our results. Not only would the number sequencing errors have to be unreasonably high, but the errors would also have to be non-randomly distributed between the populations and classes of SNPs.

In theory, some fraction of the SNPs in Table 5.2 may not be true SNPs, but instead be locations in the sequenced region with sequencing errors. If this is the case, the minor allele of these falsely identified SNPs should be at very low frequency. SNPs where the minor allele has a higher frequency are less likely to be sequencing errors, since many errors would have to have occurred at that same base. To show that our analysis is robust to this type of sequencing error, we repeated the 2 x 2 table analysis of the number of private synonymous and nonsynonymous SNPs in the AA and EA sample, excluding SNPs where the minor allele was seen only once or twice. Thus we excluded singletons and doubletons, as well as SNPs where the derived allele was almost fixed. The original 2 x 2 table (shown in Table 5.2) as well as the

modified 2 x 2 table only considering SNPs with a minor allele frequency (MAF) >2 are shown below:

All private SNPs

	S	NS	Total	% NS
AA	8958	7950	16908	0.47
EA	3879	4826	8705	0.554
	12837	12776	25613	

G = 163.2, P = 2.3 x 10^{-37}, 1 df

Private SNPs, MAF >2

	S	NS	Total	% NS
AA	2040	1503	3543	0.424
EA	807	767	1574	0.487
	2851	2275	5126	

G = 17.5, P = 2.9 x 10^{-5}, 1 df

Even limiting our analysis to SNPs with MAF > 2, we still observe a significantly higher proportion of nonsynonymous SNPs in the EA sample as compared to the AA sample. It should be pointed out that this analysis is likely highly conservative, because many weakly deleterious SNPs segregating in the population are likely to be at low frequency and thus would have been excluded from this re-analysis.

We next asked what amount of sequencing errors would be required to generate the proportional excess of nonsynonymous private SNPs in the EA population (see Table 5.2). Sequencing errors causing this pattern, in principle, could occur in two ways: 1) too few nonsynonymous SNPs could have been identified in the AA population (relative to those identified in the EA sample), or 2) some fraction of SNPs found in the EA sample are not true SNPs. Since ~75% of coding sites are nonsynonymous, most of these sequencing errors would lead to the false identification of nonsynonymous SNPs. We will consider these types of sequencing errors in turn.

To determine how many nonsynonymous SNPs would have to have been missed in the AA sample to explain the proportional excess of nonsynonymous private SNPs in the EA sample, we added nonsynonymous SNPs to the AA sample until the *P*-value for the G-test test was ~0.05. Shown below are the original 2 x 2 table for the private SNP analysis (left), the modified (now non-significant) 2 x 2 table where we

added SNPs to the boxed cell (middle), and the table showing the number of SNPs that were added to the original table to generate the modified non-significant table.

All private SNPs

	S	NS	Total
AA	8958	7950	16908
EA	3879	4826	8705
	12837	12776	25613

G = 163.2, P = 2.3 x 10^{-37}, 1 df

Modified, Non-significant table

	S	NS	Total
AA	8958	10600	19558
EA	3879	4826	8705
	12837	15426	28263

G = 3.7, P = 0.053, 1 df

Number of SNPs added

	S	NS	Total
AA	0	2650	2650
EA	0	0	0
	0	2650	2650

As can be seen in the above table ~2,650 nonsynonymous SNPs would have to have been missed in the AA sample to explain the effect we observed in the original analysis (~15%). While this amount of missing SNPs is certainly possible, the above calculations assume that no synonymous SNPs were missed in the AA sample, and that no SNPs were missed in the EA sample. If some nonsynonymous SNPs were missing from the EA sample too, then the number of missing nonsynonymous SNPs would have to be even higher, unreasonably so, in the AA sample. Next, we performed a similar analysis, this time addressing the possibility that too many SNPs were identified in the EA sample. Since changes to about 75% of coding sites would lead to a nonsynonymous SNP, about 75% of the falsely identified SNPs in the EA sample would be nonsynonymous. We dropped SNPs from the EA sample, where 75% of the dropped SNPs were from the nonsynonymous category, and 25 % of the dropped SNPs were from the synonymous category until the P-value for the G-test was ~0.05. Shown below are the original 2 x 2 table for the private SNP analysis (left), the modified (now non-significant) 2 x 2 table where we subtracted SNPs from the boxed cells (middle), and the table showing the number of SNPs that were subtracted from the original table to generate the modified non-significant table.

All private SNPs

	S	NS	Total
AA	8958	7950	16908
EA	3879	4826	8705
	12837	12776	25613

Modified, Non-significant table

	S	NS	Total
AA	8958	7950	16908
EA	3306	3107	6413
	12264	11057	23321

Number of SNPs removed

	S	NS	Total
AA	0	0	0
EA	-573	-1719	-2292
	-573	-1719	-2292

$G = 163.2$, $P = 2.3 \times 10^{-37}$, 1 df $G = 3.8$, $P = 0.051$, 1 df

As can be seen in the above table ~2,300 SNPs would have to be falsely identified SNPs in the EA sample (~ 26% of all SNPs identified in the EA sample would be false). Again, a false positive rate of ~26% of identified SNPs is not necessarily that unreasonable, but the key point is, that this false positive rate of 26% in the EA sample assumes a false positive rate of 0% in the AA sample. If any of the identified SNPs in the AA sample are not true SNPs, then the proportion of falsely identified SNPs in the EA sample would be well above 26%.

Thus, based on the above analyses, it is very unlikely that sequencing errors could solely explain the proportion excess of private nonsynonymous SNPs in the EA sample relative to the AA sample. For sequencing errors to explain our results, there would have to be a very high error rate in one population, and virtually no errors made in the other population. There is no evidence to suggest this is the case.

Supplementary Note 3

Analysis of HapMap data

The International HapMap Consortium (International HapMap Consortium 2007) genotyped ~3.8 million SNPs in 30 parent offspring trios from Utah residents with ancestry from northern and western Europe (CEU) as well as 30 parent offspring trios from the Yoruba in Ibadan, Nigeria (YRI). We wanted to determine whether these data supported our finding of an increased proportion of nonsynonymous SNPs in the European population. Additionally, the Applera dataset used for the analyses described in the main paper was based on African American individuals instead of

African individuals from Africa. The HapMap data allow us to determine whether our conclusions are robust to some level of admixture in the African American sample.

Since the number of SNPs segregating in a population is affected by how the SNPs were ascertained, we wanted to use a set of SNPs for our analysis that were discovered in a uniform manner. Therefore, we only used SNPs in the HapMap Phase II that were discovered by Perlegen Sciences re-sequencing a multi-ethnic panel (Hinds *et al.* 2005). Additionally, for a SNP to be included in our analysis, we required that there be no missing data in either the CEU or YRI sample. Finally, we also required that either human allele match both the chimp and macaque alleles. 449,797 SNPs passed these filters.

We next counted the number of synonymous and nonsynonymous SNPs that were private to either population. Functional annotations were found from dbSNP build 126. If a SNP had different annotations, it was dropped from the analysis.

Shown below is a 2 x 2 table showing the number of private synonymous and nonsynonymous SNPs segregating in YRI or CEU.

HapMap Private SNPs

	S	NS	Total	% NS
YRI	307	282	589	0.48
CEU	221	316	537	0.59
	528	598	1126	

$G = 13.6$, $P = 2.2 \times 10^{-4}$, 1 df

As can be seen in the table, the proportion of nonsynonymous SNPs is significantly higher in the CEU as compared to the YRI for the private SNPs in the HapMap dataset. Furthermore, the proportions of nonsynonymous SNPs in the HapMap dataset are similar to what we found in the Applera dataset (shown in Table 5.2). For example, we find ~48% of private SNPs in the YRI sample are nonsynonymous, compared to ~ 47% in the African American sample from the Applera dataset.

Additionally, we find ~59% of private SNPs in the CEU sample are nonsynonymous, compared to ~55.4% in the European sample in the Applera dataset.

Additionally, we reasoned that for SNPs segregating in both populations, the derived allele would be at a higher frequency in the population that underwent a bottleneck (CEU) than in the population that did not experience a bottleneck (YRI). The reason for this is that during a bottleneck, drift will have a greater effect at changing allele frequencies in the smaller population than in the larger population. We expect the frequency to be higher in the bottlenecked population because we are conditioning on the SNP segregating in both populations. Furthermore, we would expect this effect to be stronger for SNPs that are under negative selection since they will be kept at lower frequency in the larger population.

For SNPs segregating in both the YRI and CEU populations, we recorded the number of SNPs where the derived allele had a higher frequency in CEU and the number of times where the derived allele was at a higher frequency in YRI for intronic, synonymous, and nonsynonymous SNPs. The 2 x 2 tables below show the comparison of nonsynonymous and intronic SNPs (left) and the comparison of synonymous and intronic SNPs (right).

Nonsynonymous vs intronic

	NS	Intron	Ratio NS:int
Higher CEU	1100	78629	0.0140
Higher YRI	655	52429	0.0125

$G = 5.2$, $P = 0.022$, 1 df

Synonymous vs intronic

	S	Intron	Ratio S:int
Higher CEU	960	78629	0.0122
Higher YRI	622	52429	0.0119

$G = 0.31$, $P = 0.58$, 1 df

The table on the left shows that the proportion of nonsynonymous SNPs where the derived allele is at a higher frequency in CEU is higher than that expected based on intronic SNPs. To show that this is due to nonsynonymous SNPs being different, we find no difference between synonymous and intronic SNPs. When comparing

nonsynonymous to synonymous SNPs, the result is not significant, presumably due to the smaller sample size.

While these analyses of the HapMap data may be susceptible to ascertainment bias, where SNPs of a certain type are preferentially discovered and genotyped, the overall results support our finding of an increased proportion of nonsynonymous SNPs in European population relative to African populations. We tried to limit the ascertainment bias by only including SNPs that were discovered by a uniform ascertainment strategy. These analyses using the YRI individuals instead of African Americans suggest that our original finding in the Applera dataset is not due to using African American individuals instead of Africans from Africa.

REFERENCES

Adams, A. M., and R. R. Hudson, 2004 Maximum-likelihood estimation of demographic parameters using the frequency spectrum of unlinked single-nucleotide polymorphisms. Genetics **168:** 1699-1712.

Akey, J. M., M. A. Eberle, M. J. Rieder, C. S. Carlson, M. D. Shriver *et al.*, 2004 Population history and natural selection shape patterns of genetic variation in 132 genes. PLoS Biol. **2:** e286.

Anderson, E. C., and M. Slatkin, 2007 Estimation of the number of individuals founding colonized populations. Evolution **61:** 972-983.

Andolfatto, P., 2001 Adaptive hitchhiking effects on genome variability. Curr. Opin. Genet. Dev. **11:** 635-641.

Andolfatto, P., and M. Przeworski, 2001 Regions of lower crossing over harbor more rare variants in African populations of *Drosophila melanogaster*. Genetics **158:** 657-665.

Andolfatto, P., and M. Nordborg, 1998 The effect of gene conversion on intralocus associations. Genetics **148:** 1397-1399.

Aquadro, C. F., D. J. Begun and E. C. Kindahl, 1994 Selection, recombination, and DNA polymorphism in *Drosophila*, pp. 46-56 in *Non-Neutral Evolution: Theories and Molecular Data*, edited by B. Golding. Chapman & Hall, New York.

Ardlie, K., S. N. Liu-Cordero, M. A. Eberle, M. Daly, J. Barrett *et al.*, 2001 Lower-than-expected linkage disequilibrium between tightly linked markers in humans suggests a role for gene conversion. Am. J. Hum. Genet. **69:** 582-589.

Auton, A., and G. McVean, 2007 Recombination rate estimation in the presence of hotspots. Genome Res. **17:** 1219-1227.

Auton, A., K. Bryc, A. R. Boyko, K. E. Lohmueller, J. Novembre *et al.*, 2009 Global distribution of genomic diversity underscores rich complex history of continental human populations. Genome Res. **19:** 795-803.

Barbujani, G., and D. B. Goldstein, 2004 Africans and Asians abroad: Genetic diversity in Europe. Annu. Rev. Genomics Hum. Genet. **5:** 119-150.

Begun, D. J., and P. Whitley, 2000 Reduced X-linked nucleotide polymorphism in *Drosophila simulans*. Proc. Natl. Acad. Sci. U. S. A. **97:** 5960-5965.

Begun, D. J., and C. F. Aquadro, 1992 Levels of naturally occurring DNA polymorphism correlate with recombination rates in *D. melanogaster*. Nature **356:** 519-520.

Bittles, A. H., and J. V. Neel, 1994 The costs of human inbreeding and their implications for variations at the DNA level. Nat. Genet. **8:** 117-121.

Boyko, A. R., S. H. Williamson, A. R. Indap, J. D. Degenhardt, R. D. Hernandez *et al.*, 2008 Assessing the evolutionary impact of amino acid mutations in the human genome. PLoS Genet. **4:** e1000083.

Box, G. E., 1979 *Robustness in Statistics*. Academic, New York.

Braverman, J. M., R. R. Hudson, N. L. Kaplan, C. H. Langley and W. Stephan, 1995 The hitchhiking effect on the site frequency spectrum of DNA polymorphisms. Genetics **140:** 783-796.

Bryc, K., A. Auton, M. B. Nelson, J. R. Oksenberg, S. L. Hauser *et al.*, 2010 Genome-wide patterns of population structure and admixture in west Africans and African Americans. Proc. Natl. Acad. Sci. U. S. A. (in press).

Bustamante, C. D., A. Fledel-Alon, S. Williamson, R. Nielsen, M. T. Hubisz *et al.*, 2005 Natural selection on protein-coding genes in the human genome. Nature **437:** 1153-1157.

Caicedo, A. L., S. H. Williamson, R. D. Hernandez, A. Boyko, A. Fledel-Alon *et al.*, 2007 Genome-wide patterns of nucleotide polymorphism in domesticated rice. PLoS Genet. **3:** 1745-1756.

Campbell, M. C., and S. A. Tishkoff, 2008 African genetic diversity: Implications for human demographic history, modern human origins, and complex disease mapping. Annu. Rev. Genomics Hum. Genet. **9:** 403-433.

Cargill, M., D. Altshuler, J. Ireland, P. Sklar, K. Ardlie *et al.*, 1999 Characterization of single-nucleotide polymorphisms in coding regions of human genes. Nat. Genet. **22:** 231-238.

Carlson, C. S., D. J. Thomas, M. A. Eberle, J. E. Swanson, R. J. Livingston *et al.*, 2005 Genomic regions exhibiting positive selection identified from dense genotype data. Genome Res. **15:** 1553-1565.

Chakraborty, R., and K. M. Weiss, 1988 Admixture as a tool for finding linked genes and detecting that difference from allelic association between loci. Proc. Natl. Acad. Sci. U. S. A. **85:** 9119-9123.

Charlesworth, B., J. A. Coyne and N. H. Barton, 1987 The relative rates of evolution of sex chromosomes and autosomes. Am. Nat. **130:** 113-146.

Charlesworth, B., M. T. Morgan and D. Charlesworth, 1993 The effect of deleterious mutations on neutral molecular variation. Genetics **134:** 1289-1303.

Chen, G. K., P. Marjoram and J. D. Wall, 2009 Fast and flexible simulation of DNA sequence data. Genome Res. **19:** 136-142.

Chimpanzee Sequencing and Analysis Consortium, 2005 Initial sequence of the chimpanzee genome and comparison with the human genome. Nature **437:** 69-87.

Clark, A. G., 1990 Inference of haplotypes from PCR-amplified samples of diploid populations. Mol. Biol. Evol. **7:** 111-122.

Clark, A. G., M. J. Hubisz, C. D. Bustamante, S. H. Williamson and R. Nielsen, 2005 Ascertainment bias in studies of human genome-wide polymorphism. Genome Res. **15:** 1496-1502.

Clark, V. J., S. E. Ptak, I. Tiemann, Y. Qian, G. Coop *et al.*, 2007 Combining sperm typing and linkage disequilibrium analyses reveals differences in selective pressures or recombination rates across human populations. Genetics **175:** 795-804.

Cohen, J. C., R. S. Kiss, A. Pertsemlidis, Y. L. Marcel, R. McPherson *et al.*, 2004 Multiple rare alleles contribute to low plasma levels of HDL cholesterol. Science **305:** 869-872.

Collins, F. S., L. D. Brooks and A. Chakravarti, 1998 A DNA polymorphism discovery resource for research on human genetic variation. Genome Res. **8:** 1229-1231.

Conrad, D. F., M. Jakobsson, G. Coop, X. Wen, J. D. Wall *et al.*, 2006 A worldwide survey of haplotype variation and linkage disequilibrium in the human genome. Nat. Genet. **38:** 1251-1260.

Coop, G., J. K. Pickrell, J. Novembre, S. Kudaravalli, J. Li *et al.*, 2009 The role of geography in human adaptation. PLoS Genet. **5:** e1000500.

Cox, M. P., D. A. Morales, A. E. Woerner, J. Sozanski, J. D. Wall *et al.*, 2009 Autosomal resequence data reveal late stone age signals of population expansion in sub-Saharan African foraging and farming populations. PLoS One **4:** e6366.

Crawford, D. C., T. Bhangale, N. Li, G. Hellenthal, M. J. Rieder *et al.*, 2004 Evidence for substantial fine-scale variation in recombination rates across the human genome. Nat. Genet. **36:** 700-706.

Depaulis, F., and M. Veuille, 1998 Neutrality tests based on the distribution of haplotypes under an infinite-site model. Mol. Biol. Evol. **15:** 1788-1790.

Eberle, M. A., P. C. Ng, K. Kuhn, L. Zhou, D. A. Peiffer *et al.*, 2007 Power to detect risk alleles using genome-wide tag SNP panels. PLoS Genet. **3:** 1827-1837.

Evans, P. D., S. L. Gilbert, N. Mekel-Bobrov, E. J. Vallender, J. R. Anderson *et al.*, 2005 *Microcephalin*, a gene regulating brain size, continues to evolve adaptively in humans. Science **309:** 1717-1720.

Ewens, W. J., 1973 Testing for increased mutation rate for neutral alleles. Theor. Popul. Biol. **4:** 251-258.

Ewens, W. J., 1972 The sampling theory of selectively neutral alleles. Theor. Popul. Biol. **3:** 87-112.

Eyre-Walker, A., and P. D. Keightley, 1999 High genomic deleterious mutation rates in hominids. Nature **397:** 344-347.

Fagundes, N. J., N. Ray, M. Beaumont, S. Neuenschwander, F. M. Salzano *et al.*, 2007 Statistical evaluation of alternative models of human evolution. Proc. Natl. Acad. Sci. U. S. A. **104:** 17614-17619.

Falush, D., M. Stephens and J. K. Pritchard, 2003 Inference of population structure using multilocus genotype data: Linked loci and correlated allele frequencies. Genetics **164:** 1567-1587.

Fay, J. C., and C. I. Wu, 2000 Hitchhiking under positive Darwinian selection. Genetics **155:** 1405-1413.

Fay, J. C., G. J. Wyckoff and C. I. Wu, 2001 Positive and negative selection on the human genome. Genetics **158:** 1227-1234.

Fearnhead, P., and P. Donnelly, 2002 Approximate likelihood methods for estimating local recombination rate. J. R. Stat. Soc. B. **64:** 1-64.

Francois, O., M. G. Blum, M. Jakobsson and N. A. Rosenberg, 2008 Demographic history of European populations of *Arabidopsis thaliana*. PLoS Genet. **4:** e1000075.

Frisse, L., R. R. Hudson, A. Bartoszewicz, J. D. Wall, J. Donfack *et al.*, 2001 Gene conversion and different population histories may explain the contrast between polymorphism and linkage disequilibrium levels. Am. J. Hum. Genet. **69:** 831-843.

Fu, Y. X., 1995 Statistical properties of segregating sites. Theor. Popul. Biol. **48:** 172-197.

Fu, Y. X., and W. H. Li, 1993 Statistical tests of neutrality of mutations. Genetics **133:** 693-709.

Garrigan, D., and M. F. Hammer, 2006 Reconstructing human origins in the genomic era. Nat. Rev. Genet. **7:** 669-680.

Garrigan, D., Z. Mobasher, S. B. Kingan, J. A. Wilder and M. F. Hammer, 2005 Deep haplotype divergence and long-range linkage disequilibrium at xp21.1 provide evidence that humans descend from a structured ancestral population. Genetics **170:** 1849-1856.

Griffiths, R. C., and S. Tavaré, 1994 Simulating probability distributions in the coalescent. Theor. Popul. Biol **46:** 131-159.

Griffiths, R. C., and S. Tavaré, 1998 The age of a mutation in a general coalescent tree. Stoch. Models **14:** 273-295.

Gutenkunst, R. N., R. D. Hernandez, S. H. Williamson and C. D. Bustamante, 2009 Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. PLoS Genet. **5:** e1000695.

Hammer, M. F., F. L. Mendez, M. P. Cox, A. E. Woerner and J. D. Wall, 2008 Sex-biased evolutionary forces shape genomic patterns of human diversity. PLoS Genet. **4:** e1000202.

Hartl, D., and A. G. Clark, 2007 *Principles of Population Genetics*. Sinauer Associates, Sunderland, MA.

Hartl, D. L., E. N. Moriyama and S. A. Sawyer, 1994 Selection intensity for codon bias. Genetics **138:** 227-234.

Hedrick, P. W., 2007 Sex: Differences in mutation, recombination, selection, gene flow, and genetic drift. Evolution **61:** 2750-2771.

Hellenthal, G., and M. Stephens, 2007 msHOT: Modifying Hudson's ms simulator to incorporate crossover and gene conversion hotspots. Bioinformatics **23:** 520-521.

Hellenthal, G., A. Auton and D. Falush, 2008 Inferring human colonization history using a copying model. PLoS Genet. **4:** e1000078.

Hellenthal, G., J. K. Pritchard and M. Stephens, 2006 The effects of genotype-dependent recombination, and transmission asymmetry, on linkage disequilibrium. Genetics **172:** 2001-2005.

Hellmann, I., I. Ebersberger, S. E. Ptak, S. Paabo and M. Przeworski, 2003 A neutral explanation for the correlation of diversity with recombination rates in humans. Am. J. Hum. Genet. **72:** 1527-1535.

Hellmann, I., K. Prufer, H. Ji, M. C. Zody, S. Paabo *et al.*, 2005 Why do human diversity levels vary at a megabase scale? Genome Res. **15:** 1222-1231.

Hermisson, J., and P. S. Pennings, 2005 Soft sweeps: Molecular population genetics of adaptation from standing genetic variation. Genetics **169:** 2335-2352.

Hernandez, R. D., 2008 A flexible forward simulator for populations subject to selection and demography. Bioinformatics **24:** 2786-2787.

Hernandez, R. D., S. H. Williamson and C. D. Bustamante, 2007a Context dependence, ancestral misidentification, and spurious signatures of natural selection. Mol. Biol. Evol. **24:** 1792-1800.

Hernandez, R. D., M. J. Hubisz, D. A. Wheeler, D. G. Smith, B. Ferguson *et al.*, 2007b Demographic histories and patterns of linkage disequilibrium in Chinese and Indian rhesus macaques. Science **316:** 240-243.

Hinds, D. A., L. L. Stuve, G. B. Nilsen, E. Halperin, E. Eskin *et al.*, 2005 Whole-genome patterns of common DNA variation in three human populations. Science **307:** 1072-1079.

Hoggart, C. J., M. Chadeau-Hyam, T. G. Clark, R. Lampariello, J. C. Whittaker *et al.*, 2007 Sequence-level population simulations over large genomic regions. Genetics **177:** 1725-1731.

Hudson, R. R., 1993 The how and why of generating gene genealogies., pp. 23-36 in *Mechanisms of Molecular Evolution*, edited by N. Takahata and A. G. Clark. Sinauer, Sunderland, MA.

Hudson, R. R., 2002 Generating samples under a Wright-Fisher neutral model of genetic variation. Bioinformatics **18:** 337-338.

Hudson, R. R., 2001 Two-locus sampling distributions and their application. Genetics **159:** 1805-1817.

Hudson, R. R., 1983 Properties of a neutral allele model with intragenic recombination. Theor. Popul. Biol. **23:** 183-201.

Hwang, D. G., and P. Green, 2004 Bayesian Markov chain Monte Carlo sequence analysis reveals varying neutral substitution patterns in mammalian evolution. Proc. Natl. Acad. Sci. U. S. A. **101:** 13994-14001.

Innan, H., and Y. Kim, 2004 Pattern of polymorphism after strong artificial selection in a domestication event. Proc. Natl. Acad. Sci. U. S. A. **101:** 10667-10672.

Innan, H., K. Zhang, P. Marjoram, S. Tavaré and N. A. Rosenberg, 2005 Statistical tests of the coalescent model based on the haplotype frequency distribution and the number of segregating sites. Genetics **169:** 1763-1777.

International HapMap Consortium, 2007 A second generation human haplotype map of over 3.1 million SNPs. Nature **449:** 851-861.

International HapMap Consortium, 2005 A haplotype map of the human genome. Nature **437:** 1299-1320.

Jakobsson, M., S. W. Scholz, P. Scheet, J. R. Gibbs, J. M. VanLiere *et al.*, 2008 Genotype, haplotype and copy-number variation in worldwide human populations. Nature **451:** 998-1003.

Jeffreys, A. J., R. Neumann, M. Panayi, S. Myers and P. Donnelly, 2005 Human recombination hot spots hidden in regions of strong marker association. Nat. Genet. **37:** 601-606.

Jensen, J. D., Y. Kim, V. B. DuMont, C. F. Aquadro and C. D. Bustamante, 2005 Distinguishing between selective sweeps and demography using DNA polymorphism data. Genetics **170:** 1401-1410.

Kaplan, N. L., R. R. Hudson and C. H. Langley, 1989 The "hitchhiking effect" revisited. Genetics **123:** 887-899.

Karolchik, D., R. Baertsch, M. Diekhans, T. S. Furey, A. Hinrichs *et al.*, 2003 The UCSC genome browser database. Nucleic Acids Res. **31:** 51-54.

Keinan, A., J. C. Mullikin, N. Patterson and D. Reich, 2009 Accelerated genetic drift on chromosome X during the human dispersal out of Africa. Nat. Genet. **41:** 66-70.

Keinan, A., J. C. Mullikin, N. Patterson and D. Reich, 2007 Measurement of the human allele frequency spectrum demonstrates greater genetic drift in East Asians than in Europeans. Nat. Genet. **39:** 1251-1255.

Kelley, J. L., J. Madeoy, J. C. Calhoun, W. Swanson and J. M. Akey, 2006 Genomic signatures of positive selection in humans and the limits of outlier approaches. Genome Res. **16:** 980-989.

Kent, W. J., 2002 BLAT--the BLAST-like alignment tool. Genome Res. **12:** 656-664.

Kent, W. J., R. Baertsch, A. Hinrichs, W. Miller and D. Haussler, 2003 Evolution's cauldron: Duplication, deletion, and rearrangement in the mouse and human genomes. Proc. Natl. Acad. Sci. U. S. A. **100:** 11484-11489.

Kere, J., 2001 Human population genetics: Lessons from Finland. Annu. Rev. Genomics Hum. Genet. **2:** 103-128.

Kidd, J. M., Z. Cheng, T. Graves, B. Fulton, R. K. Wilson *et al.*, 2008 Haplotype sorting using human fosmid clone end-sequence pairs. Genome Res. **18:** 2016-2023.

Kim, Y., 2006 Allele frequency distribution under recurrent selective sweeps. Genetics **172:** 1967-1978.

Kim, Y., and W. Stephan, 2003 Selective sweeps in the presence of interference among partially linked loci. Genetics **164:** 389-398.

Kimura, M., 1964 Diffusion models in population genetics. J. Appl. Prob. **1:** 177-232.

Kimura, R., A. Fujimoto, K. Tokunaga and J. Ohashi, 2007 A practical genome scan for population-specific strong selective sweeps that have reached fixation. PLoS ONE **2:** e286.

Kondrashov, A. S., 1995 Contamination of the genome by very slightly deleterious mutations: Why have we not died 100 times over? J. Theor. Biol. **175:** 583-594.

Kong, A., D. F. Gudbjartsson, J. Sainz, G. M. Jonsdottir, S. A. Gudjonsson *et al.*, 2002 A high-resolution recombination map of the human genome. Nat. Genet. **31:** 241-247.

Kuhner, M. K., J. Yamato and J. Felsenstein, 1995 Estimating effective population size and mutation rate from sequence data using Metropolis-Hastings sampling. Genetics **140:** 1421-1430.

Lao, O., T. T. Lu, M. Nothnagel, O. Junge, S. Freitag-Wolf *et al.*, 2008 Correlation between genetic and geographic structure in Europe. Curr. Biol. **18:** 1241-1248.

Lautenberger, J. A., J. C. Stephens, S. J. O'Brien and M. W. Smith, 2000 Significant admixture linkage disequilibrium across 30 cM around the *FY* locus in African Americans. Am. J. Hum. Genet. **66:** 969-978.

Leblois, R., and M. Slatkin, 2007 Estimating the number of founder lineages from haplotypes of closely linked SNPs. Mol. Ecol. **16:** 2237-2245.

Lercher, M. J., and L. D. Hurst, 2002 Human SNP variability and mutation rate are higher in regions of high recombination. Trends Genet. **18:** 337-340.

Li, J. Z., D. M. Absher, H. Tang, A. M. Southwick, A. M. Casto *et al.*, 2008 Worldwide human relationships inferred from genome-wide patterns of variation. Science **319:** 1100-1104.

Li, N., and M. Stephens, 2003 Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data. Genetics **165:** 2213-2233.

Lind, J. M., H. B. Hutcheson-Dilks, S. M. Williams, J. H. Moore, M. Essex *et al.*, 2007 Elevated male European and female African contributions to the genomes of African American individuals. Hum. Genet. **120:** 713-722.

Livingston, R. J., A. von Niederhausern, A. G. Jegga, D. C. Crawford, C. S. Carlson *et al.*, 2004 Pattern of sequence variation across 213 environmental response genes. Genome Res. **14:** 1821-1831.

Lohmueller, K. E., C. D. Bustamante and A. G. Clark, 2009 Methods for human demographic inference using haplotype patterns from genomewide single-nucleotide polymorphism data. Genetics **182:** 217-231.

Lohmueller, K. E., A. R. Indap, S. Schmidt, A. R. Boyko, R. D. Hernandez *et al.*, 2008 Proportionally more deleterious genetic variation in European than in African populations. Nature **451:** 994-997.

Marchini, J., D. Cutler, N. Patterson, M. Stephens, E. Eskin *et al*, 2006 A comparison of phasing algorithms for trios and unrelated individuals. Am. J. Hum. Genet. **78:** 437-450.

Marjoram, P., and S. Tavaré, 2006 Modern computational approaches for analysing molecular genetic variation data. Nat. Rev. Genet. **7:** 759-770.

Marjoram, P., and J. D. Wall, 2006 Fast "coalescent" simulation. BMC Genet. **7:** 16.

Marth, G. T., E. Czabarka, J. Murvai and S. T. Sherry, 2004 The allele frequency spectrum in genome-wide human variation data reveals signals of differential demographic history in three large world populations. Genetics **166:** 351-372.

McVean, G., 2007 The structure of linkage disequilibrium around a selective sweep. Genetics **175:** 1395-1406.

McVean, G. A., and N. J. Cardin, 2005 Approximating the coalescent with recombination. Philos. Trans. R. Soc. Lond. B. Biol. Sci. **360:** 1387-1393.

McVean, G. A., S. R. Myers, S. Hunt, P. Deloukas, D. R. Bentley *et al.*, 2004 The fine-scale structure of recombination rate variation in the human genome. Science **304:** 581-584.

Mekel-Bobrov, N., S. L. Gilbert, P. D. Evans, E. J. Vallender, J. R. Anderson *et al.*, 2005 Ongoing adaptive evolution of *ASPM*, a brain size determinant in *Homo sapiens*. Science **309:** 1720-1722.

Morton, N. E., J. F. Crow and H. J. Muller, 1956 An estimate of the mutational damage in man from data on consanguineous marriages. Proc. Natl. Acad. Sci. U. S. A. **42:** 855-863.

Muller, H. J., 1950 Our load of mutations. Am. J. Hum. Genet. **2:** 111-176.

Myers, S., C. Fefferman and N. Patterson, 2008 Can one learn history from the allelic spectrum? Theor. Popul. Biol. **73:** 342-348.

Myers, S., L. Bottolo, C. Freeman, G. McVean and P. Donnelly, 2005 A fine-scale map of recombination rates and hotspots across the human genome. Science **310:** 321-324.

Ng, P. C., and S. Henikoff, 2006 Predicting the effects of amino acid substitutions on protein function. Annu. Rev. Genomics Hum. Genet. **7:** 61-80.

Nielsen, R., 2005 Molecular signatures of natural selection. Annu. Rev. Genet. **39:** 197-218.

Nielsen, R., 2000 Estimation of population parameters and recombination rates from single nucleotide polymorphisms. Genetics **154:** 931-942.

Nielsen, R., and M. A. Beaumont, 2009 Statistical inferences in phylogeography. Mol. Ecol. **18:** 1034-1047.

Nielsen, R., and J. Wakeley, 2001 Distinguishing migration from isolation: A Markov chain Monte Carlo approach. Genetics **158:** 885-896.

Nielsen, R., M. J. Hubisz and A. G. Clark, 2004 Reconstituting the frequency spectrum of ascertained single-nucleotide polymorphism data. Genetics **168:** 2373-2382.

Nielsen, R., I. Hellmann, M. Hubisz, C. Bustamante and A. G. Clark, 2007 Recent and ongoing selection in the human genome. Nat. Rev. Genet. **8:** 857-868.

Nielsen, R., S. Williamson, Y. Kim, M. J. Hubisz, A. G. Clark *et al.*, 2005 Genomic scans for selective sweeps using SNP data. Genome Res. **15:** 1566-1575.

Nielsen, R., M. J. Hubisz, I. Hellmann, D. Torgerson, A. M. Andres *et al.*, 2009 Darwinian and demographic forces affecting human protein coding genes. Genome Res. **19:** 838-849.

Nielsen, R., C. Bustamante, A. G. Clark, S. Glanowski, T. B. Sackton *et al.*, 2005 A scan for positively selected genes in the genomes of humans and chimpanzees. PLoS Biol. **3:** e170.

Ohta, T., 1973 Slightly deleterious mutant substitutions in evolution. Nature **246:** 96-98.

O'Reilly, P. F., E. Birney and D. J. Balding, 2008 Confounding between recombination and selection, and the Ped/Pop method for detecting selection. Genome Res. **18:** 1304-1313.

Padhukasahasram, B., P. Marjoram and M. Nordborg, 2004 Estimating the rate of gene conversion on human chromosome 21. Am. J. Hum. Genet. **75:** 386-397.

Padhukasahasram, B., J. D. Wall, P. Marjoram and M. Nordborg, 2006 Estimating recombination rates from single-nucleotide polymorphisms using summary statistics. Genetics **174:** 1517-1528.

Parra, E. J., A. Marcini, J. Akey, J. Martinson, M. A. Batzer *et al.*, 1998 Estimating African American admixture proportions by use of population-specific alleles. Am. J. Hum. Genet. **63:** 1839-1851.

Patterson, N., N. Hattangadi, B. Lane, K. E. Lohmueller, D. A. Hafler *et al.*, 2004 Methods for high-density admixture mapping of disease genes. Am. J. Hum. Genet. **74:** 979-1000.

Pawitan, Y., 2001 *In all Likelihood: Statistical Modeling and Inference using Likelihood*. Oxford University Press, Oxford.

Pennings, P. S., and J. Hermisson, 2006a Soft sweeps III: The signature of positive selection from recurrent mutation. PLoS Genet. **2:** e186.

Pennings, P. S., and J. Hermisson, 2006b Soft sweeps II--molecular population genetics of adaptation from recurrent mutation or migration. Mol. Biol. Evol. **23:** 1076-1084.

Pfaff, C. L., E. J. Parra, C. Bonilla, K. Hiester, P. M. McKeigue *et al.*, 2001 Population structure in admixed populations: Effect of admixture dynamics on the pattern of linkage disequilibrium. Am. J. Hum. Genet. **68:** 198-207.

Pickrell, J. K., G. Coop, J. Novembre, S. Kudaravalli, J. Z. Li *et al.*, 2009 Signals of recent positive selection in a worldwide sample of human populations. Genome Res. **19:** 826-837.

Plagnol, V., and J. D. Wall, 2006 Possible ancestral structure in human populations. PLoS Genet. **2:** e105.

Pluzhnikov, A., A. Di Rienzo and R. R. Hudson, 2002 Inferences about human demography based on multilocus analyses of noncoding sequences. Genetics **161:** 1209-1218.

Pool, J. E., and R. Nielsen, 2007 Population size changes reshape genomic patterns of diversity. Evolution **61:** 3001-3006.

Press, W. H., B. P. Flannery, S. A. Teukolsky and W. T. Vatterling, 1992 *Numerical Recipes in C-the Art of Scientific Computing*. Cambridge University Press, Cambridge.

Price, A. L., A. Tandon, N. Patterson, K. C. Barnes, N. Rafaels *et al.*, 2009 Sensitive detection of chromosomal segments of distinct ancestry in admixed populations. PLoS Genet. **5:** e1000519.

Pritchard, J. K., and M. Przeworski, 2001 Linkage disequilibrium in humans: Models and data. Am. J. Hum. Genet. **69:** 1-14.

Pritchard, J. K., M. Stephens and P. Donnelly, 2000 Inference of population structure using multilocus genotype data. Genetics **155:** 945-959.

Przeworski, M., 2002 The signature of positive selection at randomly chosen loci. Genetics **160:** 1179-1189.

Przeworski, M., and J. D. Wall, 2001 Why is there so little intragenic linkage disequilibrium in humans? Genet. Res. **77:** 143-151.

Przeworski, M., G. Coop and J. D. Wall, 2005 The signature of positive selection on standing genetic variation. Evolution **59:** 2312-2323.

Ptak, S. E., and M. Przeworski, 2002 Evidence for population growth in humans is confounded by fine-scale population structure. Trends Genet. **18:** 559-563.

Ptak, S. E., K. Voelpel and M. Przeworski, 2004 Insights into recombination from patterns of linkage disequilibrium in humans. Genetics **167:** 387-397.

Ptak, S. E., D. A. Hinds, K. Koehler, B. Nickel, N. Patil *et al.*, 2005 Fine-scale recombination patterns differ between chimpanzees and humans. Nat. Genet. **37:** 429-434.

Ramensky, V., P. Bork and S. Sunyaev, 2002 Human non-synonymous SNPs: Server and survey. Nucleic Acids Res. **30:** 3894-3900.

Reed, F. A., and S. A. Tishkoff, 2006 African human diversity, origins and migrations. Curr. Opin. Genet. Dev. **16:** 597-605.

Reich, D. E., S. F. Schaffner, M. J. Daly, G. McVean, J. C. Mullikin *et al.*, 2002 Human genome sequence variation and the influence of gene history, mutation and recombination. Nat. Genet. **32:** 135-142.

Reich, D. E., M. Cargill, S. Bolk, J. Ireland, P. C. Sabeti *et al.*, 2001 Linkage disequilibrium in the human genome. Nature **411:** 199-204.

Reiner, A. P., C. S. Carlson, E. Ziv, C. Iribarren, C. E. Jaquish *et al.*, 2007 Genetic ancestry, population sub-structure, and cardiovascular disease-related traits

among African-American participants in the CARDIA study. Hum. Genet. **121:** 565-575.

Romeo, S., L. A. Pennacchio, Y. Fu, E. Boerwinkle, A. Tybjaerg-Hansen *et al.*, 2007 Population-based resequencing of *ANGPTL4* uncovers variations that reduce triglycerides and increase HDL. Nat. Genet. **39:** 513-516.

Rosser, Z. H., P. Balaresque and M. A. Jobling, 2009 Gene conversion between the X chromosome and the male-specific region of the Y chromosome at a translocation hotspot. Am. J. Hum. Genet. **85:** 130-134.

Rozen, S., H. Skaletsky, J. D. Marszalek, P. J. Minx, H. S. Cordum *et al.*, 2003 Abundant gene conversion between arms of palindromes in human and ape Y chromosomes. Nature **423:** 873-876.

Sabeti, P. C., P. Varilly, B. Fry, J. Lohmueller, E. Hostetter *et al.*, 2007 Genome-wide detection and characterization of positive selection in human populations. Nature **449:** 913-918.

Sankararaman, S., G. Kimmel, E. Halperin and M. I. Jordan, 2008a On the inference of ancestries in admixed populations. Genome Res. **18:** 668-675.

Sankararaman, S., S. Sridhar, G. Kimmel and E. Halperin, 2008b Estimating local ancestry in admixed populations. Am. J. Hum. Genet. **82:** 290-303.

Santiago, E., and A. Caballero, 2005 Variation after a selective sweep in a subdivided population. Genetics **169:** 475-483.

Sawyer, S. A., and D. L. Hartl, 1992 Population genetics of polymorphism and divergence. Genetics **132:** 1161-1176

Schaffner, S. F., C. Foo, S. Gabriel, D. Reich, M. J. Daly *et al.*, 2005 Calibrating a coalescent simulation of human genome sequence variation. Genome Res. **15:** 1576-1583.

Scheet, P., and M. Stephens, 2006 A fast and flexible statistical model for large-scale population genotype data: Applications to inferring missing genotypes and haplotypic phase. Am. J. Hum. Genet. **78:** 629-644.

Seldin, M. F., T. Morii, H. E. Collins-Schramm, B. Chima, R. Kittles *et al.*, 2004 Putative ancestral origins of chromosomal segments in individual African Americans: Implications for admixture mapping. Genome Res. **14:** 1076-1084.

Simonsen, K. L., G. A. Churchill and C. F. Aquadro, 1995 Properties of statistical tests of neutrality for DNA polymorphism data. Genetics **141:** 413-429.

Singh, N. D., A. M. Larracuente and A. G. Clark, 2008 Contrasting the efficacy of selection on the X and autosomes in Drosophila. Mol. Biol. Evol. **25:** 454-467.

Skaletsky, H., T. Kuroda-Kawaguchi, P. J. Minx, H. S. Cordum, L. Hillier *et al.*, 2003 The male-specific region of the human Y chromosome is a mosaic of discrete sequence classes. Nature **423:** 825-837.

Slatkin, M., 2004 A population-genetic test of founder effects and implications for Ashkenazi Jewish diseases. Am. J. Hum. Genet. **75:** 282-293.

Slatkin, M., and R. R. Hudson, 1991 Pairwise comparisons of mitochondrial DNA sequences in stable and exponentially growing populations. Genetics **129:** 555-562.

Smith, J. M., and J. Haigh, 1974 The hitch-hiking effect of a favourable gene. Genet. Res. **23:** 23-35.

Spencer, C. C., and G. Coop, 2004 SelSim: A program to simulate population genetic data with natural selection and recombination. Bioinformatics **20:** 3673-3675.

Spencer, C. C., P. Deloukas, S. Hunt, J. Mullikin, S. Myers *et al.*, 2006 The influence of recombination on human genetic diversity. PLoS Genet. **2:** e148.

Stadler, T., B. Haubold, C. Merino, W. Stephan and P. Pfaffelhuber, 2009 The impact of sampling schemes on the site frequency spectrum in nonequilibrium subdivided populations. Genetics **182:** 205-216.

Stajich, J. E., and M. W. Hahn, 2005 Disentangling the effects of demography and selection in human history. Mol. Biol. Evol. **22:** 63-73.

Sunyaev, S., V. Ramensky, I. Koch, W. Lathe 3rd, A. S. Kondrashov *et al.*, 2001 Prediction of deleterious human alleles. Hum. Mol. Genet. **10:** 591-597.

Tajima, F., 1989a The effect of change in population size on DNA polymorphism. Genetics **123:** 597-601.

Tajima, F., 1989b Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics **123:** 585-595.

Tajima, F., 1983 Evolutionary relationship of DNA sequences in finite populations. Genetics **105:** 437-460.

Tang, H., M. Coram, P. Wang, X. Zhu and N. Risch, 2006 Reconstructing genetic ancestry blocks in admixed individuals. Am. J. Hum. Genet. **79:** 1-12.

Tang, K., K. R. Thornton and M. Stoneking, 2007 A new approach for using genome scans to detect recent positive selection in the human genome. PLoS Biol. **5:** e171.

Templeton, A. R., 2009 Statistical hypothesis testing in intraspecific phylogeography: Nested clade phylogeographical analysis vs. approximate Bayesian computation. Mol. Ecol. **18:** 319-331.

Tenesa, A., P. Navarro, B. J. Hayes, D. L. Duffy, G. M. Clarke *et al.*, 2007 Recent human effective population size estimated from linkage disequilibrium. Genome Res. **17:** 520-526.

Teshima, K. M., and M. Przeworski, 2006 Directional positive selection on an allele of arbitrary dominance. Genetics **172:** 713-718.

Teshima, K. M., G. Coop and M. Przeworski, 2006 How reliable are empirical genomic scans for selective sweeps? Genome Res. **16:** 702-712.

Thornton, K., and P. Andolfatto, 2006 Approximate Bayesian inference reveals evidence for a recent, severe bottleneck in a Netherlands population of *Drosophila melanogaster*. Genetics **172:** 1607-1619.

Thornton, K. R., and J. D. Jensen, 2007 Controlling the false-positive rate in multilocus genome scans for selection. Genetics **175:** 737-750.

Tian, C., D. A. Hinds, R. Shigeta, R. Kittles, D. G. Ballinger *et al.*, 2006 A genomewide single-nucleotide-polymorphism panel with high ancestry information for African American admixture mapping. Am. J. Hum. Genet. **79:** 640-649.

Tishkoff, S. A., F. A. Reed, F. R. Friedlaender, C. Ehret, A. Ranciaro *et al.*, 2009 The genetic structure and history of Africans and African Americans. Science **324:** 1035-1044.

Tishkoff, S. A., and S. M. Williams, 2002 Genetic analysis of African populations: Human evolution and complex disease. Nat. Rev. Genet. **3:** 611-621.

Travis, J. M., T. Munkemuller, O. J. Burton, A. Best, C. Dytham *et al.*, 2007 Deleterious mutations can surf to high densities on the wave front of an expanding population. Mol. Biol. Evol. **24:** 2334-2343.

Voight, B. F., S. Kudaravalli, X. Wen and J. K. Pritchard, 2006 A map of recent positive selection in the human genome. PLoS Biol. **4:** e72.

Voight, B. F., A. M. Adams, L. A. Frisse, Y. Qian, R. R. Hudson *et al.*, 2005 Interrogating multiple aspects of variation in a full resequencing data set to infer human population size changes. Proc. Natl. Acad. Sci. U. S. A. **102:** 18508-18513.

Wakeley, J., 2008 *Coalescent Theory: An Introduction*. Roberts and Company Publishers, Greenwood Village, CO.

Wall, J. D., 2000a A comparison of estimators of the population recombination rate. Mol. Biol. Evol. **17:** 156-163.

Wall, J. D., 2000b Detecting ancient admixture in humans using sequence polymorphism data. Genetics **154:** 1271-1279.

Wall, J. D., and J. K. Pritchard, 2003 Assessing the performance of the haplotype block model of linkage disequilibrium. Am. J. Hum. Genet. **73:** 502-515.

Wall, J. D., K. E. Lohmueller and V. Plagnol, 2009 Detecting ancient admixture and estimating demographic parameters in multiple human populations. Mol. Biol. Evol. **26:** 1823-1827.

Wall, J. D., M. P. Cox, F. L. Mendez, A. Woerner, T. Severson *et al.*, 2008 A novel DNA sequence database for analyzing human demographic history. Genome Res. **18:** 1354-1361.

Wang, E. T., G. Kodama, P. Baldi and R. K. Moyzis, 2006 Global landscape of recent inferred Darwinian selection for *Homo sapiens*. Proc. Natl. Acad. Sci. U. S. A. **103:** 135-140.

Warburton, P. E., J. Giordano, F. Cheung, Y. Gelfand and G. Benson, 2004 Inverted repeat structure of the human genome: The X-chromosome contains a preponderance of large, highly homologous inverted repeats that contain testes genes. Genome Res. **14:** 1861-1869.

Watterson, G. A., 1975 On the number of segregating sites in genetical models without recombination. Theor. Popul. Biol. **7:** 256-276.

Weiss, G., and A. von Haeseler, 1998 Inference of population history using a likelihood approach. Genetics **149:** 1539-1546.

Wigginton, J. E., D. J. Cutler and G. R. Abecasis, 2005 A note on exact tests of Hardy-Weinberg equilibrium. Am. J. Hum. Genet. **76:** 887-893.

Williamson, S. H., M. J. Hubisz, A. G. Clark, B. A. Payseur, C. D. Bustamante *et al.*, 2007 Localizing recent adaptive evolution in the human genome. PLoS Genet. **3:** e90.

Williamson, S. H., R. Hernandez, A. Fledel-Alon, L. Zhu, R. Nielsen *et al.*, 2005 Simultaneous inference of selection and population growth from patterns of variation in the human genome. Proc. Natl. Acad. Sci. U. S. A. **102:** 7882-7887.

Winckler, W., S. R. Myers, D. J. Richter, R. C. Onofrio, G. J. McDonald *et al.*, 2005 Comparison of fine-scale recombination rates in humans and chimpanzees. Science **308:** 107-111.

Wright, S., 1938 The distribution of gene frequencies under irreversible mutation. Proc. Natl. Acad. Sci. U. S. A. **24:** 253-259.

Xu, S., W. Huang, H. Wang, Y. He, Y. Wang *et al.*, 2007 Dissecting linkage disequilibrium in African-American genomes: Roles of markers and individuals. Mol. Biol. Evol. **24:** 2049-2058.

Yue, P., and J. Moult, 2006 Identification and analysis of deleterious human SNPs. J. Mol. Biol. **356:** 1263-1274.

Zeng, K., S. Mano, S. Shi and C. I. Wu, 2007 Comparisons of site- and haplotype-frequency methods for detecting positive selection. Mol. Biol. Evol. **24:** 1562-1574.

Zhou, Y. H., and W. H. Li, 1996 Gene conversion and natural selection in the evolution of X-linked color vision genes in higher primates. Mol. Biol. Evol. **13:** 780-783.

Zhu, L., and C. D. Bustamante, 2005 A composite-likelihood approach for detecting directional selection from DNA sequence data. Genetics **170:** 1411-1421.